Dae ye ken me?: Speech Synthesis in the Gorbals Region of Glasgow

Alexis Grant

Supervisors: Simon King, Rob Clark, and Korin Richmond

Master of Science
in
Speech and Language Processing
Theoretical and Applied Linguistics
School of Philosophy, Psychology and Language Sciences
University of Edinburgh

2005

Abstract

This report attempts to determine whether improving the phonetic match between a lexicon used in speech synthesis and the speech of the speaker who provides the source for a synthetic voice improves the quality and authenticity of the synthetic voice. In order to explore this question, a survey is made of the literature describing the accent of urban Glasgow, and these data serve as the basis for the accent's implementation in the Unisyn accent-independent lexicon system. After implementation, two voices are built from the accent, one using an automatic labeling procedure and one using a corrected labeling. The voices are assessed subjectively for quality and compared in forced-choice listening tests to reference voices built using the Edinburgh accent of the Unisyn lexicon. The results are inconclusive, but the method is generalizable and potentially powerful.

Acknowledgements

I have been very fortunate to have the direct help of many people in completing this work. First and greatest thanks go to disability office for their assistance, which allowed me to continue to work while recovering from repetitive strain injury, and to Angela Chitaznidi and Tim Mills, who realized that assistance by serving as my hands through many long and sometimes tedious hours of lab work.

Mark Fraser made my life far easier by doing the initial segmentation and transcription of the Gorbals recordings, and providing me with comparison voices based on the tailored Edinburgh lexicon. Volker Strom assisted by providing the basis for the script that creates the listening test. Cassie Mayo provided references for understanding speech perception tests.

I am very grateful to my supervisors for their help. Simon King was particularly supportive of my medical needs throughout the process, and Rob Clark provided much-needed technical assistance with Festival and statistics.

Very big thanks go to all the people who participated in my listening experiments, kindly donating their time and auditory acuity to furthering the cause of science.

I have also received a great deal of indirect support in my pursuit of this goal. My parents have always encouraged and supported me in my endeavors, and provided a great deal of practical and emotional support during this work. My friends' positivity and belief in me was invaluable. I am very lucky to include among my friends Mark, whose presence in my life reminds me that the best things can happen when we least expect them, and Adam, without whose support I would be considerably less healthy and happy than I am today.

Finally, I am greatly indebted to all the medical professionals who helped me on my road to recovery. I would especially like to thank Jenny Tyler, physiotherapist and Pilates instructor, and Penny Wall, massage therapist and acupuncturist. Their consistent belief in my eventual recovery and completion of this work demonstrates both great professional dedication and great personal warmth.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text. This work has not been submitted for any other degree or professional qualification except as specified.

(*Alexis Grant*)

To Nancy Niedzielski, who inspired me to pursue my interest in linguistics

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Speech synthesis is an area of research that has many applications in the real world. A synthesized voice can make repeated announcements that would bore a human and wear out his voice, but is more flexible than a recorded inventory of preset phrases. In recent years, speech synthesis has improved in quality with advances in the algorithms used to select, dissect, and reassemble a database of recorded speech into new utterances. Instead of small databases of diphone units, with only one example of each diphone, modern synthesizers are able to select from large databases of recorded speech that contain multiple units of a given phonetic type. The natural variation found in such large databases allows the synthesizer to select units that may already be similar to the desired output units. By reducing the need for signal processing, the quality and naturalness of the output speech is greatly improved. This process, called unit selection synthesis, can also reduce the number of joins between units and check adjacent units for good matches, reducing discontinuity (Hunt and Black, 1996).

However, these new large single speaker databases are generally still dependent on the same lexicons that synthetic voices have always been based on. These lexicons are large collections of words, each word transcribed phonetically into a single sequence of phones. These transcriptions attempt to provide a reasonably accurate and realistic pronunciation of the words, but can only ever provide one or a few pronunciations for a given word, with different pronunciations generally accompanying different parts of speech or different semantics, shown by a part of speech or semantic identifier in the lexicon, and carefully disambiguated by the speech synthesis system.

The pronunciations in such lexicons are generally based on the dominant accent of the country in which the lexicon was developed: normally General American English in the

United States or Received Pronunciation (RP) in Britain. They are normally used not only in the process of generating synthesis output, but also in the process of labeling the speech databases.

1.1 The Accent Problem

Because large databases of speech may be involved, it is usual to use an automatic process to label the speech utterances in preparation for the process of dividing them into different units. If these lexicons are used to label the utterances of a speaker with a significantly different accent from the one represented in the lexicon, the labels that are generated will necessarily be incorrect. Sounds with different phonetic realizations may receive the same label; some sounds that are present in the labels will not be present in the speech, and vice versa. Similar problems will occur when dictionary lookup during synthesis produces a phone sequence and suggests units to choose.

1.2 Possible Solutions

One possible solution would be to continue to confine voice development to those with similar accents to the lexicons already developed. At first glance, this may seem to be a reasonable solution—those accents are widely intelligible throughout the countries in which they're used and are linguistically well-documented. It is therefore possible to obtain a high-quality voice that many people will understand. However, the business of producing a voice that speaks a particular accent is not solely affected by quality and intelligibility. Any accent gives a form of social identity to the voice: in addition to its being young or old, male or female, it suggests membership in a social subgroup that may be defined by many factors, including ethnicity, social class, or geographical location. In Britain, the RP accent is stereotypically associated with native English people of the middle or upper classes. For some applications, a synthesizer with this type of affiliation might not be the most appropriate. A good example would be a telephone synthesis system for a Scottish bank (Williams and Isard, 1997). It is possible that the bank would prefer to display an image that is more Scottish, especially to its Scottish customers, who might feel more comfortable with a voice that appears to be "one of them". Similarly, the relevant application for the current work is a conversational speech synthesis system that will form part of an art installation in the Gorbals area of urban Glasgow. One goal of this installation is to document the

local language and culture. To do this, the voices of local people will be required, and attempting to fit their voices to an American or English accent is neither feasible nor suitable for the purpose. It is clear from these examples that confining voice development to accents currently implemented in lexicons is not a good solution.

A second solution would be to develop a new lexicon for each accent from scratch. It does not take very much reflection to see that this would be incredibly time-consuming and inefficient. The differences between accents are often slight in quality, but they affect many words, which in this scenario would at a minimum have to be individually copied from a developed lexicon and edited; they might even have to be re-created.

The obvious improvement to a solution involving the creation of a new lexicon would be to convert currently existing lexicons to new accents through various rules. This would result in many complex rules dealing with insertions, deletions, and transformations, many of which would have to interrelate the spelling and pronunciation, but it is nevertheless the most feasible solution discussed thus far. Unfortunately, this solution runs afoul of the very large problem: no single accent contains all the distinctions made in other accents. People in most parts of Britain pronounce the vowel in the word 'bruise' similarly to that in the word 'goose', so the common RP lexicons contain the same phonetic symbol for these vowels. But people in Wales pronounce these two vowels differently. Sometimes the distinction can be elucidated by the spelling, but sometimes it cannot, so this solution would still involve a large amount of tedious hand editing (Fitt, 1997).

A third possible solution, which will be the one used in this study, is to use a metalexicon that represents a pronunciation with all its possible distinctions and uses a transformation mechanism to produce output for various accents. The Unisyn lexicon attempts to address this problem in an economical way by providing such a lexicon and transformation mechanism (Fitt, 2000). Pronunciations in the Unisyn lexicon are not represented by phonetic sequences, but instead are represented using keysymbols: symbols that consistently behave as a group across accents. Each keysymbol is based on a word, called the keyword, that represents a group of words with the relevant symbol. For example, the word FOOT represents a set of words with the keysymbol |u|, realized in both General American English and RP as the vowel /ʊ/. The realization may be different in other accents, but it will be the same throughout the set. The conversion of keysymbols into phonetic sequences is achieved by a set of rules and mappings. The keysymbol approach is based on the keyvowel approach of Wells (1982)

but is expanded to cover consonants and include variations that Wells didn't consider. A number of accents have been implemented in this lexicon, including the Scottish accents of Edinburgh and Aberdeen.

1.3 Goals and Methods of the Present Study

The goal of the present study is to implement a third Scottish accent for the Unisyn lexicon: the accent of urban Glasgow. I will exploit published information about the accent to create phonetic rules in the lexicon system that will produce sociolinguistically accurate output. This requires adding resources to the lexicon to account for a phenomenon that is particularly pertinent to this accent – that of realization variation. While everybody pronounces words and sentences differently at different times, Glaswegian has several phonetic rules that only operate a certain percentage of the time, and any attempt to capture the accent authentically must take this into account. The implemented accent will be tested by building a synthetic voice using the accent for labeling and output purposes. The speech database for the voice is an older male speaker from a spontaneous speech corpus recorded in the Gorbals during 2004.

Evaluation of the voice will involve comparison with a synthetic voice built for the Festival speech synthesis system (Black et al., 1999) from the same speech database but using the Edinburgh accent implemented for the Unisyn lexicon. The voice built using the Edinburgh accent also uses certain additional words and pronunciations added to account for some of the obvious features of the dialect. The effect of the accent on the labeling procedure will be evaluated by examining some outcomes from the automatic labeling, and the overall quality of the voice will be given a subjective evaluation with particular emphasis on labeling issues. Two speech perception listening tests will examine whether the new accent has improved the naturalness or authenticity of the voice.

Chapter 2

Unisyn Lexicon

The system used by the Unisyn lexicon to represent various accents in one metalexicon bears further examination in order to understand how the Glaswegian accent will be implemented in the lexicon. The format of the lexicon includes the orthography of the word, an identifier (often numeric, sometimes containing semantic information) if two words of the same orthography have different pronunciations, the word's part of speech, a pronunciation field, and enriched orthography field, and a frequency field. The fields are separated by single colons. Shown below is a typical entry:

```
abated::VBN/VBD: { @ . b * ee t }.> I7 d > :{abate}>ed>:427
```

This entry does not contain an identifier; if one were present it would be between the two colons separating the orthography from the part of speech. The slash between the two parts of speech indicates that the word could be either. If instead a vertical bar were present, it would indicate that the word is simultaneously both parts of speech (for example, contractions are simultaneously modals and adverbs). The most important field is the pronunciation field. The |@| represents a schwa class, the basic unstressed vowel. The |b| is also straightforward, as are the other consonants. The |ee| is represented by the word WASTE, one of two vowels realized in RP as /ɛɪ/ and General American as /eɪ/. The distinction exists for the purpose of some Welsh accents, and is a good example of the kind of distinction that would not normally be represented in a mainstream lexicon. Following the pronunciation, the enriched orthography shows morpheme boundaries and allows the lexicon to identify related words by their common morphemes. The frequency field derives from a composite of online sources of word frequency, including the British National Corpus. It is not currently used in creating rules, but could conceivably play a role in applying rules selectively only to

highly frequent items. This gives an idea of the potential power and flexibility of the system.

2.1 Keysymbols

The use of keysymbols rather than phonetic symbols is the heart of the Unisyn system. Whenever a keysymbol appears in the lexicon, it always represents the same underlying pronunciation as other occurrences of the keysymbol. However, it may represent different pronunciations for different accents. This concept originates with Wells (1982), which used keyvowels based on General American and RP accents to discuss accent variation. The Unisyn lexicon extends Wells's method to cover consonants as well, because some variation between accents occurs in consonants, although most of it is concentrated in vowels. In Wells's work, each keyvowel symbol is represented by a word that contains that symbol.

For example, the symbol |ou| is represented by the keyword GOAT, and any words containing that vowel are referred to as words of the GOAT class. In General American, the pronunciation of that vowel is /oʷ/, whereas in RP, the pronunciation is /eʊ/. But both can be represented in the same way in the lexicon, because any vowel in that class will be pronounced in the same way. This means that separate lexicons are not required to represent the separate pronunciations. This is an important move toward unifying lexicons representing a variety of different accents and reducing the work required to create a new one: it eliminates the need for two entirely separate pronunciations. The creation of separate pronunciations is replaced by the one-time creation of the metalexicon by determining appropriate keysymbol pronunciations for words, followed by a mapping between a keysymbol and its phonetic equivalent for each accent being represented. This is only the first step toward the creation of a true metalexicon. Much accent variation occurs because the distribution and phonemic classification of sounds is different between accents, rather than just phonetic quality. Some of this variation is allophonic, and is taken care of in the Unisyn lexicon by transformations that convert one keysymbol to another based on the phonetic environment (Fitt and Isard, 1998). The integration of lexicon and transformation system makes the Unisyn lexicon really more of a lexicon system than just a lexicon. The system consists of the metalexicon plus the many transformations that deal with differences in phonetic distribution. In order to facilitate the formulation of complex transformations, the Unisyn lexicon makes use of many additional marks representing syllable, morpheme, and word boundaries,

as well as lexical stress. Some of these marks can be seen in the entry cited above: braces to surround a free morpheme, angle brackets to surround a bound morpheme, a period to represent syllable boundaries, and an asterisk to represent primary stress. Thus, complex transformations dependent on things like syllable boundary or word-final position (such as the occurrence of glottal stops in some British accents) can be represented in the lexicon system and applied before producing final transcriptions for an accent. There are distinct keysymbols for allophones, which only appear in output transcriptions.

Parallel to allophones, which only appear in output transcripts, there are some keysymbols that only appear in the lexicon itself. For example, there are several keysymbols that represent vowels that are kept as full vowels in some accents, but reduced in others. There is no need for these symbols to appear in output transcripts, because they're always transformed to other keysymbols by rule: either the reduced vowel or the full vowel of the appropriate class. One of these symbols is used in the following excerpt from the list of Unisyn keywords.

|I6| PIRATE { p * ae . r I6 t } |I6| → |i|, |@| conv_i_schwa_6

In this excerpt, the keyword PIRATE represents the keysymbol |I6|, which is either reduced or retained as |i|. Before output, the symbol is identified by a rule called `conv_i_schwa_6`. If the setting for the accent indicates that the vowel should be reduced, it is converted to the keysymbol |@|; otherwise, it is converted to |i|. This alternative then appears in the output transcript. There are also keysymbols that represent subclasses of vowels that are realized as different full vowels in different accents. One symbol in this class is the |ah2| vowel, which is represented as a subclass of the BATH vowel |ah|. It is pronounced as a low front vowel in some accents, such as Australian English, and as a low back vowel in accents such as RP. This vowel is a separate keysymbol, but, like the full/reduced vowels, is not produced in output transcripts because in each accent it has the same behavior as a class that has already been established. In Australian English it merges with TRAP, whereas in RP it merges with BATH. These examples show that some differences in keysymbol behavior can't be specified through a transformation whose environment is defined by phonetic and typographical markers. Differences like these must be defined inside the metalexicon by using a different keysymbol, which can then be identified easily by rules. Most variation of this type involves differences created by accent change over time. However, sometimes allophonic variation or other phonetic processes like reduction occur in environments

that cannot be identified by Unisyn transformations, as in the example above, and thus necessitate new keysymbols.

2.2 Producing Accent-Specific Text

The process of producing accent-specific text from the lexicon can be used to produce an accent-specific lexicon, giving pronunciations specific to the accent using appropriate keysymbols. But it can also produce transcriptions for running text. Because the lexicon includes rules that work across word boundaries, using it to transcribe running text gives a more accurate transcription compared to simply concatenating words from an accent-specific lexicon. Using the system to transcribe the classic text "hello world" gives the following results for some of the main accents:

```
RP:               #{ h @ . l * ou }#.#{ w * @@r lw d }#
Edinburgh:        #{ h @ . l * ou }#.#{ w * @@r r l d }#
General American: #{ h @ . l * ou }#.#{ w * @@r r lw d }#
```

This gives the transcription for the complete utterance, not just for each word individually. It is easy to see that there are several keysymbol differences between these three accents. One difference is that RP and General American use the symbol |lw| before the final consonant, indicating that there is some difference from the standard behavior of the keysymbol |l|, which is the keysymbol present in the lexicon for this word. Notice that in RP, a non-rhotic accent, the |r| is missing. However, the previous vowel retains an |r| in its keysymbol, signaling that it was once pre-rhotic. This is a convention used in the lexicon to simplify the specification of various rules pertaining to rhoticity. In the other accents, the |r| is retained. There may be further phonetic differences that correspond to different pronunciations of keysymbols and so are not represented at this level.

2.2.1 Hierarchy

In order to efficiently specify the behavior of different accents, the lexicon divides the accents into accent groups, resulting in a hierarchy including information about country, region, and town. For example, the Edinburgh accent is in the country of the UK, the region of Scotland, and the town of Edinburgh. Transformations that determine how the accent behaves can be written to control any of these levels. They are divided into

three types: conversions, rules, and mappings. Conversions are generally context-free and deal with single symbols, whereas rules tend to be sensitive to phonetic context and may involve more complex transformations. They are otherwise similar, and later discussion will not make the distinction. Mappings occur at the end of the transcription process and collapse classes of keysymbols whose behavior is identical in the output accent. In all cases, the transformation's behavior will depend on the score set for that rule for that accent, which is looked up during transcription. Conversions are assumed to apply to every accent, with the default conversion behavior being a score of 1; non-default behaviors are specified with scores greater than one. Rules are assumed to apply only if they have a nonzero score for that accent. Mappings can only have a score of 0 or 1, where zero means that it does not apply (default). In the case of rules and conversions, if the score is different from the default, different scores will evoke different behavior. Accent scores can then be easily set, and the design of a new accent is simplified because the bulk of the specification simply involves setting the appropriate scores for the various rules. In some cases, including, as we will see, Glasgow Vernacular, new accents require new rules and other modifications.

For example, many UK accents are non-rhotic, so the default value for UK accents in the area of rhoticity is non-rhotic with intrusive /r/ and linking /r/, but different values of the rule can be specified for regions and towns to change this. For example, RP does not have intrusive /r/, so it has a different score. For Scotland and Ireland, the rhoticity value is overridden at the regional level to make them rhotic (Fitt, 2000). Behaviors set at the regional level can be further overridden at the level of town. This allows a great flexibility in specifying accent behavior, while still being efficient because values set at levels higher in the hierarchy are inherited. In addition to these specifications based on geographical area, it is also possible to specify an individual person whose accent behaves differently from that of their general town. The hierarchy is an important aspect of the lexicon because it reflects both the general unity of accent found in geographical area and the possibility of variation within each level. However, if a rule operates for a particular accent, it always operates for that accent, and cannot vary randomly or depending on situation or interlocutor.

2.2.2 Exceptions

The lexicon also has a small list of exceptions which are kept in a separate file. Although most variation is represented within the lexicon itself, there are some variations

which would be cumbersome to represent, and so these are represented separately and merged with the lexicon before use. Frequently these are variations in stress or complex variations in pronunciation which only apply in a small region. (Some variations in pronunciation, such as consistent differences between UK accents and US accents, are represented by typographical markers in the lexicon, but this strategy is not efficient for complex variations in small regions.) Exceptions are generally applied to a basic form and all its derived forms, except when the derived form is itself an exception. Forms are matched by the enriched orthography field, which, as discussed above, captures the morphemic structure of the word. For example, in the excerpt shown above for the word 'abated', if an exception affected the root word 'abate', then it would apply to 'abated' because the root 'abate' is found in its enriched orthography field.

This rich structure provides fertile ground for accent implementation. The linguistic specification of the accent in the upcoming chapter will be integrated into this structure in Chapter 4.

Chapter 3

Glasgow Vernacular

3.1 Glasgow's Linguistic Situation

The term Glasgow Vernacular originates with Jane Stuart-Smith, who has done a great deal of contemporary research into Glasgow speech. It is used to describe the colloquial language of the working classes of Glasgow (Stuart-Smith, 1999a). Glasgow Vernacular is a distinct linguistic variety, more a dialect than an accent: it has phonological, morphological, syntactic, and lexical differences from other varieties of English, including other Scottish varieties. It is often described as a mixture or "compromise" between the remnants of the Scots language, modern Scottish English (more specifically Scottish or Glasgow Standard English), and current developments indicate that it takes additional influence from prestigious nonstandard dialects of English, such as London English (Stuart-Smith, 2003, p. 110). Standard dialects of English English, such as RP, have little prestige in Glasgow and probably do not influence its language greatly (Stuart-Smith, 1999a).

Historically, the Scots language was spoken in southern, central, and northeastern Scotland, but it has been in continuous contact with English as the prestige language variety for several hundred years. This contact has caused erosion, and much of the Scots language has fallen out of use, with the remaining parts under pressure to conform to the English standard. Macafee observes that the erosion often takes the form of what was once commonplace in universal becoming colorful and idiosyncratic, the unique features of the dialect becoming only embellishments on the daily standard (Macafee, 1994). For many language users in Scotland, Scots forms have almost entirely disappeared, with only a few common words and pronunciations remaining in use, along

with a distinctive Scottish accent. However, many language users, including those of interest for this project, continue to alternate between more traditional dialect forms and more modern English forms on all levels, preserving syntactic, lexical, and phonetic variations originating in the Scots language. Stuart-Smith writes:

> Within Wells' framework for describing English accents, an accent in a traditional dialect area can be expected to show differences from the standard most clearly in lexical incidence, but also in terms of phonemic system, phonotactic structure, and allophonic realization. (Stuart-Smith, 1999b, p. 185)

It is this situation that led to A. J. Aitken's famous model of modern Scottish speech using the "bipolar stylistic continuum", with vernacular Scots at one end and Scottish English at the other (Aitken, 1984, p. 519f). Stuart-Smith notes prominently and repeatedly that it is not clear that any present speakers can be placed purely at the Scots extreme. Rather, Glasgow Vernacular speakers are generally somewhere in the middle of that continuum, mixing standard and dialectal forms in everyday speech.The situation is further complicated by the fact that many speakers of this mixture alter their placement on the continuum depending on the context and activity. Level of traditional dialect use decreases during reading relative to speaking, and changes also occur based on context and interlocutor, with speakers being particularly prone to shift toward the English end of the continuum when a speaker outside the speech community is present.

A further complication is present in the accent variations within Glasgow itself. Although many of these variations are based on class, some Glaswegians insist that they are also based on neighborhood areas. Macaulay interviewed his subjects in this area, and found that only four of the 36 adult and 15-year-old subjects believed that it was possible to distinguish Glaswegians from different areas, and often qualified this belief by saying it was only for certain parts of the city or that they could only tell when the person was from the part of the city that they themselves were from. Others compared areas of significantly different social class, so it was not clear if they were really talking about areas or simply using them as indexes of class (Macaulay and Trevelyan, 1977, p. 87). It may be that this was more possible before the social upheaval dating from the post-World War II housing crisis. Before the redevelopment of Glasgow's inner-city areas, many neighborhoods contained tight social and family networks, but the resettlement of many people on peripheral housing estates and change in the housing situation, from tenements to more modern flat blocks, disrupted the situation.

I will focus on describing the phonological characteristics of Glasgow Vernacular, with some discussion of morphological and lexical differences that affect phonological processes and phonetic distribution. For a more general description of modern Scots, see Jones (2002).

3.2 Scottish English

An overview of some of the general characteristics of Scottish English will be helpful to establish the grounds of the discussion. Wells (1982) offers the basis for such an overview in an appropriate framework, since his work is the inspiration for the Unisyn keyword lexicon used in this work. Wells makes reference to lexical sets of words, very similar to Unisyn keyword classes.

Several notable characteristics of Scottish English include:

- There is no lax version of the high back vowel /u/, and the phonetic location of this vowel is more central (/ʉ/). In Wells's terminology, words in the FOOT class have the same vowel as words in the GOOSE class.
- There is generally only one low vowel, rather than the front and back low vowels. This amounts to homophony of the word classes TRAP, BATH, and PALM.
- The realization of the GOAT and FACE word classes are monophthongs, as compared to (different) diphthongal realizations in both General American and RP. The realizations in Scottish English are somewhat similar to those of American English, but without the diphthongal upgliding, and can be generally described as /e/ and /o/.
- LOT, CLOTH, and THOUGHT generally all have the same vowel.
- PRICE words are usually divided into two classes with two different vowels. This is associated with the length differences caused by the Scottish Vowel Length Rule (SVLR), which also applies to the vowels /i/ and /u/, and lengthens vowels in the environment of a following voiced fricative, an /r/, or a morpheme boundary. (See Scobbie et al. (1999) for more detailed description; this simple version is the one used in the Unisyn lexicon.) In the case of the PRICE vowel, it also causes a quality change, so it is simpler to have it represented in the lexicon by two separate vowels.

- Scottish English is generally rhotic, and vowels before post-consonantal /r/ may be distinguished (thus *bird* may have a different vowel from *work*, and *earth* is almost always different, except in upper-class Scottish Standard English). The realization of /r/ is traditionally a tap or retroflex approximant, although the central approximant usual in American English and RP (in non-postvocalic environments) is becoming common. The older Scots trill is rarely found in modern speakers.
- The Scottish consonant inventory includes the fricative /x/, used in dialect words such as 'loch', and the voiceless labiovelar fricative /ʍ/, used word-initially in WH words.
- In many varieties, glottal stops are commonly used as a substitute for /t/ in certain environments, and sometimes for /p/ and /k/.
- No h-dropping occurs, unlike many northern English dialects.

3.3 Characteristics of the Vernacular

The above list of characteristics concentrates on the characteristics common to most Scottish accents. Studies focused more specifically on Glasgow give evidence of other characteristics relevant to this variety.

3.3.1 Studies

3.3.1.1 Macaulay and Trevelyan

Macaulay and Trevelyan (1977) studied five "phonological variables": (i), representing the vowel in words like *bit*; (u), representing the vowel in words like *goose*; (a), representing the vowel in words like *hat*; (au), representing the vowel in words like *now*; and (gs), representing the use of glottal stop as an alternative to /t/. Macaulay used the Labovian sociolinguistic method in this investigation, dividing each variable into several possible categories along a continuum and giving each a score (Labov, as described in Macaulay and Trevelyan 1977). The speech was drawn from interviews with Glaswegian informants from three age groups (adults, teenagers, and children) who were indexed into four social classes, denoted I (professional and managerial), IIa (intermediate nonmanual), IIb (skilled manual), and III (semi and unskilled manual).

The class of teenagers and children is based on the father's occupation. Macaulay's objective was to make a comparison between the various classes, and indeed he did find differences based on age, sex, and class. This was one of the first substantial studies focused on Glasgow to give these kinds of results, and considering these differences has proved useful in later studies.

For my purposes it is primarily the results of classes IIb and III that are likely to apply to speakers of the vernacular. Macaulay's results for these two classes are more similar than different for most variables.

3.3.1.2 Stuart-Smith

Stuart-Smith has made a number of recent studies of Glasgow Vernacular. Most are based on phonetic data from recordings made in 1997 of 32 speakers, grouped by class (middle class or working class, roughly equivalent to Macaulay's class I versus his class IIb and class III), by age (younger, 13 to 14 years old; older, 40 to 60 years old) and gender (male or female) (Stuart-Smith and Tweedie, 2000). She has investigated both read speech and conversational speech. I will focus on the results for working-class conversational speech, since those results will be most relevant to the synthesis for this project. The spontaneous speech was elicited in self-selected same-sex conversational pairs, talking for around 45 minutes.

One of the major results of this study is the discussion of a large number of consonant variables, some of which involve different allophones: realization of /θ/ and /ð/, vocalization of /l/ and /r/, and loss of /x/ and /ʍ/ (Timmins et al., 2004). I will not address the issue of /l/-vocalization because Stuart-Smith reports a very small percentage of it.

3.3.1.3 Macafee

Macafee (1994) recorded 62 working-class speakers, both male and female and in a variety of age groups, from age 10 to 66+. Most of the recordings were group interviews including the investigator. The goal of the study was to investigate usage of Scots lexical items, but she also investigated some phonological variables and their lexical incidence. Her subjects were primarily from four areas in the East End of Glasgow: the Calton, Bridgeton, Dennistoun, and Barrowfield.

3.3.1.4 Johnston

Johnston (1997) has compiled an extensive discussion of regional variation in Scots, with particular attention to vowel classes and vowel alternations. His main source is the *Linguistic Atlas of Scotland* (Mather and Speitel, 1986), but he has done extensive regularization in order to present an understandable picture of vowel classes. He places Glasgow in the mid-Scots group and describes potential vowel alternations in this group, as well as discussing some consonantal processes.

3.3.2 Consonants

3.3.2.1 Glottal Stop

Use of the glottal stop is one of the most salient characteristics of Glasgow Vernacular. It is often considered a categorical use, especially where it replaces /t/, but studies show that this is not really the case. Macaulay and Trevelyan (1977)'s results for (gs) showed that the percentages for class IIb and class III were very high, at around 84% and 90% respectively. However, these are not 100%, indicating that it is important to take into account the fact that not all /t/s are glottalized even for speakers of the vernacular. Macaulay also notes that there are differences in the occurrence before a vowel or pause versus before a consonant. The occurrence is highest before a consonant, occurring in almost 100% of cases, whereas it is lowest before pauses, with vowels intermediate. Likewise, glottal stops are less likely to be used in the middle of the word than at the end, with only 68.8% of word-medial stops being glottalized by class III speakers, compared to 91.8% of word-final stops.

In Timmins et al. (2004)'s study, working-class adults produce about 90% glottal stops, very similar to Macaulay's results, with younger speakers producing on average 94%, due to a very high occurrence in adolescent females of 99%. Results for relative frequency of glottaling in various phonetic environments were generally similar to Macaulay's, except that glottal stops were more likely before a pause than word-finally before a vowel. Most of the exceptions for prevocalic glottaling occurred in sequences of two /t/s, like "put it" or "at all" (Stuart-Smith, 1999b, p. 194). She suggests that this may result from resyllabification of the /t/, and notes that avoidance of glottalization in this environment is common in other accents. Intervocalically, uses of /t/ seemed to indicate style-drifting, with speakers often starting out using /t/, or switching to it when doing imitations or reading labels. Such trends are even more

extreme for younger speakers, with variation away from the glottal stop tending to be exclusively emphatic or a result of style-drifting.

3.3.2.2 Realization of /θ/ and /ð/

Both of these phonemes have traditional variants: /h/ for /θ/ and /ɾ/ for /θ/. For younger speakers, it is now possible to substitute either of these phonemes with their labiodental equivalents, /f/ and /v/, which is classically observed as a feature of the London Cockney accent (Wells, 1982). Macafee (1994) also mentions this trend in younger speakers, and Timmins et al. (2004) describe it in detail. Older speakers do not show this change for either phoneme. For /θ/, they show a high level of use of the standard, coupled with a low level of the traditional variant, from 12% to 16% in spontaneous speech. For younger speakers, the use of the traditional variant is between 41% and 44%, with the remainder being approximately evenly divided among the new /f/ and the standard /θ/ (females show more /f/ by 10%, males vice versa). The use of the different variants varies slightly by phonetic environment, with /f/ being most common in word-initial environments, and /h/ most common in word-medial environments. Word-final environments mostly show /θ/, with a small amount of /f/. /ð/ is a somewhat more complicated case, since it mostly occurs word-initially, where it shows little variation. Word-medially, some taps were observed, with some deletion also emerging in young working-class speakers. Unlike most cases, where reading inhibits dialectal features, there is more variation in read speech than in spontaneous speech. /v/ does not occur as a substitute for /ð/ in spontaneous speech, perhaps not entirely surprising since /ð/ is not a particularly common sound except in extremely common words like "they" and "their", which might be less susceptible to such a change.

3.3.2.3 Loss of /x/ and /ʍ/

I stated in my general description of Scottish English that /x/ and /ʍ/ were part of the consonant inventory. This continues to be true, but their occurrence has been dwindling and is frequently only found in place or person names. Macafee (1994) suggests that this may be due to the influence of Hiberno-English imported with the many Irish immigrants to Glasgow, which does not possess these phonemes. In Timmins et al. (2004), spontaneous speech from older working-class people used /x/ almost categorically, but younger speakers primarily used /k/, with almost no pure /x/ being heard (some intermediate variants were observed). This is not a common phone and the

number of occurrences was quite small. For /ʍ/ loss, the results were more mixed, with working-class adolescents having lost /ʍ/ between 60 and 70% of the time, while older working-class were more split, with older females using about 82% /ʍ/ and older males about 45%. A number of intermediate realizations were heard for both phonemes.

3.3.2.4 /r/ Vocalization

The vocalization of postvocalic /r/, resulting in a change in the status of rhoticity in Glaswegian, has been the subject of extensive analysis by Jane Stuart-Smith and her colleagues as well as meriting a mention by Johnston (Johnston, 1997; Stuart-Smith, 2003; Timmins et al., 2004). They have found that there is a large variety of realizations of postvocalic /r/, moving from the traditional tap articulation through various approximants and even full vocalization. This appears to be part of a wider trend in Scots; Johnston extends it as a rule for the entire mid-Scots area except South-west Mid (Ulster), although without detail, and Romaine observed a similar variety in Edinburgh school children (Johnston, 1997; Romaine, 1978). The proportion of each articulation used, like the other variables previously described, varies depending on the age and gender of the speaker. In Stuart-Smith (2003), the highest percentage of vocalization is found in young females, with nearly 60% of postvocalic /r/ being vocalized. In contrast, older males only show about 12% vocalization. The most common phonetic /r/ variants are central approximants, retroflex approximants, and taps. The traditional tap ranges between 20 and 28%, with older speakers often using retroflex approximants but occasionally using central ones, whereas younger females prefer central approximants. Younger males tend to use retroflex approximants and taps, using hardly any central approximants.

Stuart-Smith's postvocalic data is also further subdivided into pre-consonantal and pre-pausal, where pre-pausal means occurring at the end of the word. This is further divided into /r/ occurring after stressed versus unstressed syllables, which are then again subdivided by utterance position into turn-final, prevocalic, and pre-consonantal. The alternation occurs with varying frequency in these different environments, occurring most frequently pre-consonantally, followed by a high occurrence in turn final position, particularly when the previous syllable was unstressed. This is followed by slightly lower occurrences word finally before a consonant, and least likely word finally before a vowel. However, like the pronunciation of /r/ when it does occur, vocalization in-

volves variation: it is sometimes accompanied by velarization, which patterns slightly differently from /r/-loss which produces a plain vowel.

3.3.2.5 Scots

There are also Scots processes, mostly historical, that affect consonants. Most of these, like most of the vowel alternations, originate in the different development of words in Scots versus English. These include loss of some postvocalic /l/, such as in the word *ball* or *all* (part of a change in Older Scots, a different process than the current /l/ vocalization). There has been some loss of final /d/ after /n/ (e.g. 'staun', stand), and less commonly after /l/, though Johnston (1997) mentions that this is becoming increasingly common over the mid-Scots region. There is also commonly devoicing of the past tense suffix after liquids or nasals ('pullt' for pulled). Macafee (1994) discusses a number of these alternations briefly, including the use, common across many dialects, of an -in suffix rather than an -ing suffix. /v/ was frequently vocalized in older Scots as well, resulting in modern forms like 'gie' (/gɪ/) for "give". 'Wi' for "with" is also common, with the sound also lost in words such as 'clothes' ('claes').

Macafee also helpfully describes the usual occurrence of the enclitic negative '-nae' with verb forms. This results in forms such as 'havnae', 'willnae', etc. This is a common feature of the vernacular, and was used by most of Macafee's informants, except for three who used only a small number of Scots forms. She also found that 'dinnae' for 'don't' was much less common than most '-nae' forms. She observes that one common use of 'don't' is in 'don't know', but nobody says "dinnae know"; one girl says "dinnae ken", which is certainly a common collocation in Scots vernacular generally (Macafee, 1994, p. 224). The standard English form is normally used in tag questions.

3.3.3 Vowels

Although the consonant alternations, particularly glottal stop alternations, are a very important part of the vernacular, vowel differences are perhaps even more important. Both phonetic variability and vowel class alternations are involved.

3.3.3.1 Macaulay's Phonetic Variability

Macaulay and Trevelyan (1977) studied three vowel variables, (i), (u), and (a), where the differences between the classes are primarily phonetic rather than phonological, with classes IIb and III preferring more back variants of (a) and more central variants of (i) and (u) rather than more phonetically peripheral types (high front for (i), high back for (u)), with (i) also involving a degree of retraction. For (i), something over 60% of the variants for those classes were the most retracted three variants. This is a frequently noted characteristic of Glaswegian vernacular, often parodied in comedy and the media. "Little Wullie [Willie]" as a character in stories is a common example of this kind of parody. Johnston writes:

> All Mid- and Southern Scots dialects transfer the vowel to CUT in *wind* and *pill*. The West Mid group, and neighboring West Lothian and Stirlingshire, add *hill* and, increasingly, *girl* to this transfer list, and tendencies to do so are increasing over time, spreading eastwards and southwards. (Johnston, 1997, p. 470; the italicized words are subclasses of the vowel, generally involving words with similar phonological structure.)

Macaulay and Trevelyan (1977) indicates that the phonetic environment seems to influence the variation of (i), with more front consonants seeming to correlate with higher variants of (i), whereas velars, and, oddly, dentals seem to influence it to be lower.

Macaulay's results for (u) have been criticized because he included all words that fall into the /u/ category in Scottish English, but in the vernacular some of these fall into the vowel category /ɪ/ (e.g. foot, 'fit') and some into the category /ʌ/ (e.g. pull), although in Macaulay (1978) he indicates that a recalculation of the indices for some speakers suggest that the pattern is still valid when this effect is taken into account. The issue of differing vowel class membership is a complex one and is discussed more extensively in 3.3.3.2. The variable (a) also has vowel class-related issues, with Macaulay noting that many previous writers have observed a varying distribution of phonetic low vowels in different words, but he does not seem to have explored this issue.

3.3.3.2 Vowel Alternation

Another large part of the vernacular is the incidence of different vowels in various subsets of words. This poses a large problem for an attempt to characterize the dialect using the Unisyn lexicon. The vowel classes in the Unisyn lexicon are fundamentally based on General American and RP. Although the use of vowel classes liberates a lexi-

con, and therefore the synthesizer, from the strictures of a single accent, the formation of these classes from particular accents means that a certain bias remains: an assumption that the sets of words that behaved together in the source accents will continue to behave together in other accents. A large amount of editing must go on in order to either reassign words to different vowel classes, or create new vowel classes or subclasses for a given accent.

The reason for this incidence of different vowels is the origin of the dialect in the Scots language, which had separate historical developments from English up until the 18th century. Johnston discusses the evolution of Scots vowels in his article about regional variation in Scots (Johnston, 1997). Scottish accents can be divided into a number of regions. Johnston's classification defines each accent region, explaining which areas are the center of each accent group and how accent change tends to spread throughout each group.

The classification deals only with the areas of Scotland that were Scots speaking rather than Gaelic speaking, because the Gaelic-speaking areas transitioned directly from Gaelic to English, and in consequence have a different linguistic development. So Johnston's regions encompass the South and Central parts of Scotland, as well as the Northeast and the Orkney and Shetland Islands. Glasgow, the largest city in Scotland, is the center of an important accent group and often the center of change. The Glasgow-Edinburgh urban central belt is referred to in this classification as "mid-Scots" and is divided further into central regions for the cities, several regions surrounding Glasgow, and several regions surrounding Edinburgh.

3.3.3.3 Vowel Classes in Glasgow Vernacular

Johnston (1997) discusses contemporary reflexes for each vowel class in each dialect region, with vowel classes being based on the Older Scots vowels. The difference between the Older Scots-based vowel classes and Wells (1982)'s modern RP/General American classes is clear at a glance. Where Wells has the keyword FLEECE, Johnston has two keywords, MEET and BEAT, and where Wells has FACE, Johnston again has two classes, MATE and BAIT. Word membership in these classes is also different: the MATE class includes words like *load*, *rope*, *soap*, *apple*, and *toe*, along with words with what we usually think of as the MATE vowel, like *rate*, *later*, *baby*, *safe*, and *save*.

The effect of this is that Glasgow has several major vowel alternations and a number of minor ones. Stuart-Smith (2003) shows an effective representation of these alter-

nations. This chart is reproduced in Figure 3.1 as an aid for the discussion of the alternations. One notable quality of this chart is that vowels in Scottish Standard English mostly correspond directly to RP vowels. This means that Scottish Standard English differs primarily phonetically from RP, and would thus be easy to implement in the lexicon (indeed, the Edinburgh accent for the lexicon is approximately this accent). The one exception to this is the distinction between the BITE vowel and the TRY vowel, which simply subdivides one of Wells's vowel classes, the PRICE class. However, the distribution of alternating vowels is quite different. Although both MEET and BEAT are pronounced the same, a subclass of the BEAT words may be pronounced with /e/, and, as noted, the MATE class includes words that in Scottish Standard English are pronounced with /o/. Other salient alternations include the pronunciation of some words of the GOOSE class with /i/, and some with /e/. These are noted with BOOT and DO respectively.

The most salient Scottish alternation involves words of the MOUTH class, here denoted by OUT, which are pronounced with /ʉ/ rather than /o^{w}/. It is worth emphasizing that this class does not include all words of the MOUTH class; not all of them are susceptible to this alternation. In addition to that, there are a number of differences relating primarily to the low vowels and /o/, where many Scots speakers have fewer vowels as well as different class membership. These are the vowels beginning with COT and ending with SNOW. As I mentioned above, there is generally a merger of BATH, TRAP, and PALM, as well as LOT, CLOTH, and THOUGHT, and some members of the latter in Wells's system are members of the former in Glasgow Vernacular. Johnston (1997) indicates that there may be a merger in Glasgow Vernacular between the LOT/CLOTH/THOUGHT vowel and the GOAT vowel, but it is not clear whether this is the case. Macafee (1994) notably mentions that forms like 'strang' for "strong" are susceptible to transfer to forms with /o/, because of a frequent transfer between /o/ and /ɔ/.

The alternations occur for many speakers, but no speaker has been demonstrated who consistently shows the Scots version of alternation for every case (Stuart-Smith, 2003). Although the alternating vowels are represented as belonging to classes, the class membership has been consistently eroded by contact with English, and pressure to conform to English as the prestige variety, so the class membership may vary for individuals, individuals may be aware of both class memberships and use both in different circumstances. Because of this, speakers may vary in whether they use the alternated vowel for particular lexical item, and with what frequency if so. Because of the difficulty of

Table 6.1 An outline of Scottish English vowels (after Aitken, 1984, Table 1, Macafee Chapter 7, Figure 7.2, Johnston, 1997, 453). ↔ indicates vowel alternation.

keyword	no.	Scots (Glasgow)		SSE	RP
MEET	2	i	i	i	i
BEAT	3	i	i	i	i
(DEAD)		i	i ↔ ɛ	ɛ	ɛ
MATE	4	e	e	e	eɪ
(BOTH)		e	e ↔ o	o	əʊ
BAIT	8	e	e	e	eɪ
PAY	8a	əi	əi ↔ e	e	eɪ
BOOT	7	ɪ	ɪ ↔ ʉ	ʉ	u
DO		e	e ↔ ʉ	ʉ	u
BIT	15	ɪ	ɪ	ɪ	ɪ
BET	16	ɛ	ɛ	ɛ	ɛ
OUT	6	ʉ	ʉ ↔ ʌʉ	ʌʉ	aʊ
COAT	5	o	o	o	əʊ
COT	18	o	o ↔ ɔ	ɔ	ɒ
OFF		a	a ↔ ɔ	ɔ	ɒ
CAT	17	a	a	a	a
(LONG)		a	a ↔ ɔ	ɔ	ɒ
(WASH)		a	a ↔ ɔ	ɔ	ɒ
HAND		ɔ	ɔ ↔ a	a	a
START		ɛ	ɛ ↔ a	a	a
CAUGHT	12	ɔ	ɔ	ɔ	ɔ
(SNOW)		ɔ	ɔ ↔ o	o	əʊ
CUT	19	ʌ	ʌ	ʌ	ʌ
(PULL)	6a	ʌ	ʌ ↔ ʉ	ʉ	ʊ
NEW/DEW	14	ju	ju	ju	ju
BITE	1*s*	əi	əi	əi	aɪ
TRY	1*l*	ae	ae	ae	aɪ
EYE	11	i	i ↔ ae	ae	aɪ
LOIN	10	əi	əi ↔ oe	oe	ɔɪ
VOICE	9	oe	oe	oe	ɔɪ
LOUP 'jump'	13	ʌʉ	ʌʉ	(ʌʉ)	-

Figure 3.1: Scots vowel alternations (Stuart-Smith, 2003, p. 116)

collecting large amounts of data involving these alternations, it is not even clear how many lexical items are involved anymore.

3.3.3.3.1 The MOUTH Vowel This is true even for the best studied contrast of the MOUTH vowel. Stuart-Smith (2003) notes that there are six common items, one fairly common item, and several rare items from Macafee's study, plus a few more alternations that she found. I will discuss these further in Section 4.2.5.

Macaulay and Trevelyan (1977) also studied this particular alternation. The variable (au) quantifies the distribution and various allophones of the (au) diphthong, along with the Scots alternate of a monophthong. Macaulay mentions that he has been criticized for discussing this variable as a continuum; Romaine, for example, suggests that the monophthong is a choice distinct from the diphthong (Romaine, as cited in Macaulay and Trevelyan 1977), and the preceding discussion would seem to support that: it isn't truly meaningful to discuss the overall level of monophthong use without considering the fact that some items in this class are not susceptible to this alternation. However, his treatment is informative, since it does distinguish the monophthong from the diphthong without collapsing all the variants of the diphthong together. His diphthong variants go from a realization of the diphthong similar to the RP or General American realization, /aʊ/, to more centralized realizations on both ends, /əu/, although /ʌʉ/ is now the more usual representation. He found that the use of the monophthong was very high among class III, but lower among class IIb, which prefers centralized realizations of the diphthong. This suggests that a distinct treatment is needed for the monophthongs and diphthong, and that the diphthong, among speakers in the vernacular, generally has a fairly centralized phonetic realization, although it may vary somewhat among individuals. Diphthongs are particularly prone to variance in unstressed positions, which may contribute to the variety in the data.

3.3.4 The Importance of Variation

One of the most notable features of the descriptions in this section is that almost all of them describe a percentage-based alternation that differs depending on the sociolinguistic group membership of the speaker. Sometimes, it appears that there is a change in progress. This leads to a very important point about modern Glasgow Vernacular: it is importantly characterized by variation. This variation is both within and between speakers. Macaulay writes: "The inconsistency with which forms are used has led com-

mentators to despair." (Macaulay and Trevelyan, 1977, p. 25) But we cannot despair. The fact that these alternations clearly encode social differences makes them extremely important in creating an accurate replication of the vernacular for the Unisyn lexicon and potentially for a speech synthesizer. It is not only important to be able to produce the alternations, but also to be able to alter them depending on the voice being created in order to give the voice an accurate social identity. Fortunately, the Unisyn lexicon is amenable to this type of work, although additional capacities had to be added. I will now describe how I was able to update the lexicon to take this complex situation into account.

Chapter 4

Glaswegian Vernacular in the Unisyn Lexicon

The Unisyn lexicon, as I described it above, takes in a single string of text and an instruction about what accent to use (Fitt, 2000). The main part of the Unisyn lexicon that transforms the text is called the post-lexical rules. Previous to the post-lexical rules, the string of text is turned into a string of initial pronunciations, drawn from the exceptions and the accent-independent lexicon. The system then runs a series of rules on the initial pronunciations, and produces as output a single string of pronunciations. The rules are implemented as regular-expression substitutions, finding every instance of their target pattern in the string and replacing it by the desired result. For example, one fairly simple rule works as follows:

Name: `do_ur_or`

Purpose: convert the symbol |ur| to |our| (FORCE, score 2) or |or| (NORTH, score 1) if the accent merges it with one of those two symbols. The conversion occurs before a consonant or word or compound boundary. It applies to words such as *poor* and *touring*.

Formulation: substitute |ur r| followed by a consonant or a word or compound boundary **with** |or r| or |our r| followed by the same thing that originally followed |ur r|. Optionally, other types of boundaries may occur between the vowel and the |r|, or after the |r|.

This is obviously not adequate for producing alternations, since the original pattern will always be replaced by the same thing.

4.1 Accounting for Variation

Previously, a string simply went through transformations and came out the other end. Any transformation was either applied, if it was relevant to the accent, or not applied, if it was not. I chose to replace this single result with a list of all the possible results. Now, a rule that applies to the accent may apply to all, no, or some cases of the environment in which it applies. In the first case, the initial environment is always substituted with the result; in the second case, the initial environment remains unchanged and the rule appears not to apply; the third case is somewhere in between. This reflects the real situation in Glasgow Vernacular where speakers choose whether or not to use a certain vowel or allophone in a certain context. In order to produce a list as a result, the string processed by the post-lexical rules was first transformed into the sole member of a list, which could then be added to. Post-lexical rules were modified to accept a list as input and produce one as output, keeping track of any additions made along the way.

Rules whose operation is still simple all-or-nothing, which make up the bulk of the rules in the lexicon, operate much like they did before. They simply run their standard text transformation on every item currently in the list. However, if a rule takes variation into account, it generates additional alternatives, putting more items in the list. Each alternative is given a score that represents its likelihood of being chosen by a speaker. Scores are computed as products of the scores of each chosen transformation, as usual for likelihoods. Transformation scores can be assigned for an entire rule, or only for certain environments within a rule. Each score is a probability, but multiplied by 100 so that it is greater than 1, avoiding the problem of progressively smaller results of multiplication. The transformation scores are stored in a separate file, and are loaded when the lexicon is first run and looked up at the time of making the transformation, much like the scores for the rules. The scores assigned to various rules are discussed in more detail in Section 4.2.

The process for putting the list through a rule that takes variation into account is as follows:

1. The list is fed into a rule of this type.
2. The first item in the list is removed from the list for operation.
3. The rule checks to see if any of the patterns that the rule operates on are present in the string.

4. (a) **If the pattern is present**, the rule creates as many new strings as there are possible alternations in that position. (Usually there are only two: either the original pattern is retained, or the rule applies and the result pattern replaces the original pattern. Sometimes a rule might have two or more possible result patterns, and in this case there would be three or more new strings. Sometimes the original pattern is never retained, but there are two possible result patterns, and in this case there would be two strings.)

   (b) **If the pattern is not present**, the string is returned unchanged to the list. The list is transferred out of the rule, any occurrences of the marked original string (see Step 5) are restored to their original state, and the list moves on to the next rule.

5. Each new string is transformed. Where the pattern occurs, it is substituted with the different possible result patterns. If one of the possibilities is to retain the original pattern, then the original pattern is replaced by a pattern that is identical to the original except that it is marked (currently by a trailing 1, since that is a pattern that doesn't occur with other meanings) so that it is clear that that occurrence of the pattern has been checked.

6. Each new string is entered into a table which keeps track of the relative scores of the strings. The new score of each string is its current score multiplied by the score of that result of that transformation.

7. The original item pulled off the list is removed from the table.

8. The new transformed strings are added to the end of the list.

9. The rule is called again on the new list.

Note that this is a recursive process. Step 4b is the termination condition; after all of the original occurrences of the pattern in the original and modified strings have been replaced either by marked original strings or by result patterns, the pattern will no longer be found in any of the strings. The order of the strings in the list ensures that the rule will not find a string that does not contain the original pattern (termination condition) while there remain strings containing the original pattern in the list. The marking of the original patterns that have been checked is what prevents the rule from running forever, constantly changing the original patterns for the results.

This ensures that every occurrence of the pattern in every string is checked and strings created that have each of the possibilities in that position. The net result is a long list

of strings that contains every possible way in which the string could be pronounced. When the list reaches the end of all the rules, the score of each string in the list is the product of the score of each possibility that was taken in that string. These are then sorted, so that the list is printed out in the order of highest score to lowest score.

4.1.1 Pruning

I have also implemented a mechanism for pruning, since the number of possibilities can grow very large for long strings. This works rather like hypothesis pruning in speech recognition, where a possible pronunciation is pruned if its score falls too far below the current high score. Because scores for long strings can get very large, reaching orders of magnitude around 10^{15}, an absolute threshold of score difference, where the smaller score is subtracted from the larger and the result is compared to a fixed number, will not work. Scores in the early stages may differ by only small amounts such as 100 or 200, but in the final stage scores may differ from each other by $7 * 10^{15}$. Because of this, for any given threshold almost no pruning will be done early on, and extremely aggressive pruning will run at the end, producing only one or a few possibilities. An absolute threshold is also ruled out by the fact that different sentences will have different scores, so a level of pruning that would work for a shorter sentence will not work for longer sentences. Instead, the pruning is implemented using a magnitude relative to that of the current high score. For example, if a sentence's score is three orders of magnitude smaller than the current high score, that pronunciation will be dropped from the list. The order of magnitude for pruning can be set on the command line; it defaults to 'off' and can be set back to that value at any time.

Pruning occurs primarily between rules, so that the table that has been built up during the rule is pruned down before beginning the next rule. At the end of the rule, the sentences are sorted by score, and any sentence with a score below the threshold is removed from the possibilities. However, some rules create a very large number of alternations during the time that they run. For these rules, pruning occurs during the rule, when the new scores are computed for the different alternatives generated during the current iteration (Step 5 in the replacement process, above). If the score will be too low compared to the current high score, as established at the beginning of the replacement of that pattern, then the alternative is never added. As I describe the rules, I will state whether pruning occurs during a given rule.

4.2 New Rules

4.2.1 Glottal Stop

A complex glottal stop rule already existed in the lexicon, with a score of 1 for Scottish accents. For this rule, glottal stops are used in place of /t/ in many environments: after vowels, /n/, /l/, or /r/ when the /t/ precedes a word or compound boundary (or a pause), or is syllable final before an unstressed vowel or consonant; when /t/ begins a syllable that does not begin a free morpheme, and is followed by an unstressed weak vowel; when /t/ precedes a syllabic consonant or is in a final cluster (Fitt, 2000). Although it seems likely that there is some alternation in glottal stops in other regions of Scotland, the main goal is to implement the alternation for Glasgow, so I created a new rule score that applied only to the town of Glasgow, with a score of 3. This operates in the same environments as the original rule, because the pattern of glottalization in Glasgow is similar to that in other places. However, this rule takes variation into account. Whenever it encounters an environment in the string which would normally provoke the substitution of /t/ with a glottal stop, it will create two new strings with each alternative and assign each the appropriate score. The score for the glottal stop rule is always high, but varies depending on the environment, based on my information about relative frequency from Macaulay and Stuart-Smith (Section 3.3.2.1). The scores are shown in Table 4.1.

Environment	Score
Before a weak unstressed vowel	8
Syllable-final	8.5
Before a boundary	9
In final clusters/before syllabic consonants	7

Table 4.1: Scores for glottal stop rule environments

Although there are some subtleties of glottal stop use in Glasgow (Stuart-Smith, 1999b), I judged this implementation to be sufficient to broadly reflect the variation involved. Several examples and their relative scores (scores only computed for the different glottal stop alternatives, and not for any other potential variation) are shown in Table 4.2. Relative scores will be used in all tables of this form.

Because this rule operates quite frequently and can generate many alternatives, pruning occurs during this rule if pruning is on.

Candidate	Score
#{ g * o ? }#.#{ * e . n ii }#.#{ b * uh . ? n! }> z >#	63
#{ g * o ? }#.#{ * e . n ii }#.#{ b * uh . t n! }> z >#	27
#{ g * o t }#.#{ * e . n ii }#.#{ b * uh . ? n! }> z >#	7
#{ g * o t }#.#{ * e . n ii }#.#{ b * uh . t n! }> z >#	3

Table 4.2: Unilex transcriptions of "got any buttons"

4.2.2 R-Vocalization

Previously, there was a rule in the lexicon which dealt with the conversion of the standard approximant |r| (/ɹ/) to the Scottish |tˆ| (/ɾ/). For the Edinburgh accent, this converted any non-postvocalic /r/ that was not in a consonant cluster into a tap. However, given the discussion in Section 3.3.2.4 of the complex realization of /r/ in Glaswegian, this is obviously inadequate. Thus I created from scratch an entirely new rule dealing with /r/ for Glaswegian, with a score of 4. For postvocalic /r/, this uses the same environments that were dealt with in the analysis by Stuart-Smith (2003); for intervocalic and word-initial /r/, it introduces probabilities between /ɾ/ and /ɹ/. In each case of postvocalic /r/, three possibilities are used: no /r/, /ɾ/, or /ɹ/. I made no attempt to distinguish between the two common types of approximants, retroflex (/ɻ/) and central (/ɹ/). I chose not to distinguish because the two seemed to show similar patterns of realization by phonetic context, differing only by their overall frequency in a given speaker, based on his or her sociolinguistic group membership. This means that it is acceptable for them to be selected interchangeably, and should lead to a reproduction in the synthesis of the frequency from the speaker. However, /ɾ/ is quite different in sound, and also occurs much more frequently in intervocalic and word initial cases, so it is necessary to be able to distinguish these. Similarly, the behavior of plain vowels is very different, as is their sound. Although Stuart-Smith's data showed the use of pharyngealized vowel variants, and showed that the distribution of pharyngealized versus plain vowels was different in different phonetic contexts, I judged it as too difficult to implement because the difference is difficult to perceive and label, especially with a small amount of data. It is not normal for synthesis to take into account such subtle differences in phonetic labeling for voice creation or synthesis.

Thus for all postvocalic cases of /r/, three new strings are created, one with each alternative. Each is assigned the appropriate score. Scores are given in the lexicon system as the score for the likelihood of /ɾ/ and the score for the likelihood of some

form of /r/ . Since the total likelihood is 10, the system computes the likelihood of vocalization by subtracting the likelihood of some form of /r/ from 10. The actual likelihood of an /ɹ/ is computed by subtracting the likelihood of /ɾ/ from the likelihood of /r/. For intervocalic or word-initial /r/, only the two possibilities are considered, because vocalization is not possible. The scores, shown as likelihoods in Table 4.3, vary a great deal, because the probability of use varies a great deal. Pruning occurs during this rule, because of the frequency of its application.

Environment	/ɾ/ Score	/ɹ/ Score	Vocal. Score
Intervocalic	8	2	n/a
Word-initial	9	1	n/a
Stressed word-final intervocalic	6	2	2
Unstressed word-final intervocalic	3	3	4
Stressed word-final pre-consonantal	4	3	3
Unstressed word-final pre-consonantal	2	4	4
Pre-consonantal (same word)	3	3	4
Stressed pre-pausal	3	3	4
Unstressed pre-pausal	4	3	3

Table 4.3: Scores for /r/ rule environments

Giving an example of this rule is quite complex, but the output shown in Tables 4.4 and 4.5 should give a good idea of the way the rule works.

Candidate	Score
#{ tˆ* ai ? }#.#{ k i l . b * uhr . n ii }#.#{ s t r * ii ? }#	32
#{ tˆ* ai ? }#.#{ k i l . b * uhr tˆ. n ii }#.#{ s t r * ii ? }#	24
#{ tˆ* ai ? }#.#{ k i l . b * uhr r . n ii }#.#{ s t r * ii ? }#	24
#{ r * ai ? }#.#{ k i l . b * uhr . n ii }#.#{ s t r * ii ? }#	8
#{ r * ai ? }#.#{ k i l . b * uhr tˆ. n ii }#.#{ s t r * ii ? }#	6
#{ r * ai ? }#.#{ k i l . b * uhr r . n ii }#.#{ s t r * ii ? }#	6

Table 4.4: Unilex transcriptions of "right kilburney street"

The example in Table 4.4, "right Kilburney Street", shows examples of word-initial and pre-consonantal environments. The example in Table 4.5 shows unstressed prevocalic and pre-pausal environments. In both cases, some sentences have the same score, which is normal. It simply means that the sentences are equally likely according to the system.

Candidate	Score
#{ b * uh . ? @r r }#.#{ @ n d }#.#{ sh * uu . g @r tˆ}#	24
#{ b * uh . ? @r }##{ @ n d }#.#{ sh * uu . g @r tˆ}#	16
#{ b * uh . ? @r tˆ}#.#{ @ n d }#.#{ sh * uu . g @r tˆ}#	12
#{ b * uh . ? @r }##{ @ n d }#.#{ sh * uu . g @r }#	12
#{ b * uh . ? @r r }#.#{ @ n d }#.#{ sh * uu . g @r r }#	9
#{ b * uh . ? @r r }#.#{ @ n d }#.#{ sh * uu . g @r }#	9
#{ b * uh . ? @r tˆ}#.#{ @ n d }#.#{ sh * uu . g @r }#	9
#{ b * uh . ? @r tˆ}#.#{ @ n d }#.#{ sh * uu . g @r r }#	9
#{ b * uh . ? @r }##{ @ n d }#.#{ sh * uu . g @r r }#	9

Table 4.5: Unilex transcriptions of "butter and sugar"

4.2.3 /f/ for /θ/

This alternation in theory applies to both unvoiced and voiced dental fricatives, but Timmins et al. (2004) did not record any use of the voiced fricative /v/ in spontaneous speech, so I have only implemented this alternation for the unvoiced fricatives /f/ and /θ/. The data indicates that this is primarily used by younger speakers, and that its frequency varies depending on position in the word: word-initial, word-medial, and word-final positions have different frequencies of conversion. Because of the way word boundaries are shown in Unisyn, it is not difficult to write this rule. The rule score is 1, because no other accent with this rule currently exists. In each case, two alternatives will be created with the appropriate scores. An example is shown in Table 4.6. This only shows the word initial and word medial environments, with scores of 3 and 1 respectively (so they are not too likely to occur; the standard variants are more common).

I did not define a rule that would substitute /h/ for /θ/, because it was unclear from the data whether this was lexically restricted or not.

4.2.4 /ɾ/ for /ð/

The traditional tap for /ð/ in word-medial position sometimes occurs (Johnston, 1997), and is implemented much like the previous rule, but without word-initial or word-final environments. This is shown in Table 4.7.

Candidate	Score
#{ th r * ii; }#.#{ m * a th s }#.#{ s t y * uu . d n ? }> s >#	63
#{ th r * ii; }#.#{ m * a f s }#.#{ s t y * uu . d n ? }> s >#	27
#{ f r * ii; }#.#{ m * a th s }#.#{ s t y * uu . d n ? }> s >#	7
#{ f r * ii; }#.#{ m * a f s }#.#{ s t y * uu . d n ? }> s >#	3

Table 4.6: Unilex transcriptions of "three maths students"

Candidate	Score
#{ b r * uh . dh @r r }#	8
#{ b r * uh . tˆ@r r }#	2

Table 4.7: Unilex transcriptions of "brother"

This is quite a simple rule and has a score of 2 for the likelihood.

4.2.5 MOUTH **Vowel**

I chose to implement this alternation by creating a new vowel class. The new keysymbol is |ouu|, which was chosen to maintain the association with the 'o' from the |ow| keysymbol, while adding an association with 'u'. Any word meant to be subject to the alternation was placed in this new vowel class, including words such as *about*, *our*, *round*, *down*, *out*, *now*, and *house*. These words are the seven items commonly involved in this alternation, found by Macafee (1994). Between 10 and 15 others are found infrequently, including *trousers*, *mouth*, *Southside*, and *pound* (Stuart-Smith, 2003). The rule itself is a very simple one, which is actually implemented as a conversion since it is a single symbol and is context-free. The search pattern identifies any occurrence of |ouu|. If the score for the rule is such that it doesn't apply, |ouu| is converted to |ow| for each string in the list. If the rule does apply, two new strings are created, one containing |ow| and one containing |uu|, so this is a rule where the original symbol is never maintained. I decided to do this since no data was available about whether words with the potential alternation to the monophthong tended to have noticeably different phonetic realizations from others, so they are all categorized together under |ow|. As above, each new string is assigned an appropriate score based on which alternative was chosen in that string, and the likelihood of the alternation (Table 4.8).

Candidate	Score
#{ * uu ? }#.#{ @ n d }#.#{ @ . b * uu ? }#	36
#{ * uu ? }#.#{ @ n d }#.#{ @ . b * ow ? }#	24
#{ * ow ? }#.#{ @ n d }#.#{ @ . b * uu ? }#	24
#{ * ow ? }#.#{ @ n d }#.#{ @ . b * ow ? }#	16

Table 4.8: Unilex transcriptions of "out and about"

The score for this alternation is 6, which is roughly the frequency with which the highly frequent words alternate (Stuart-Smith, 2003).

4.2.6 Standard /k/ and /w/ for /x/ and /ʍ/

This is similar to the MOUTH vowel, but there isn't a new keysymbol. Instead, a simple conversion finds any occurrence of /x/ and creates two new strings, one containing /x/ and one containing /k/ (score of 6). The same thing occurs for /ʍ/ and /w/ (score of 5), and these are scored appropriately (Table 4.9).

Candidate	Score
#{ hw * eir tˆ}#.#{ i z }#.#{ dh @ }#.#{ l * o x }#	30
#{ w * eir tˆ}#.#{ i z }#.#{ dh @ }#.#{ l * o x }#	30
#{ hw * eir tˆ}#.#{ i z }#.#{ dh @ }#.#{ l * o k }#	20
#{ w * eir tˆ}#.#{ i z }#.#{ dh @ }#.#{ l * o k }#	20

Table 4.9: Unilex transcriptions of "where is the loch"

4.2.7 Alternations

The alternations in general are dealt with in a slightly different way. Although the alternations do occur in classes of words, the vowel that occurs in Scots may not be phonetically close to to the vowel that occurs in Scottish English or other dialects, so representing them as subclasses of current vowel classes would be infeasible. Also, there are very many of them, and the creation of so many new keysymbols for vowels that are phonetically the same as vowels that are already in the keysymbol lexicon is not a solution that is economical or easy to maintain. The orthography might also be challenging, since the vowel should relate closely to the vowel it's realizing, while

potentially also being orthographically linked to the vowel it alternates with. This could lead to symbol combinations that do not make obvious sense. Arguably, these are synchronically simply erratic alternations, because the classes are so small, but conceptually they actually arise from a completely different set of vowel classes as found in older Scots. Thus, representing them as subclasses of current classes would not really capture either the synchronic or diachronic reality.

Instead, I have chosen to follow the model already instantiated in the lexicon in the form of unpredictable American versus British pronunciation differences. These are shown with a forward-slash, where the vowel in one dialect is simply different from the vowel in the other. The best example is the word 'tomato', which is pronounced as a back vowel in RP and other British accents, but a fairly high front vowel in the US. This is represented as follows:

```
tomato::NN: { t @ . m * aa/ee . t ou } :{tomato}:2723
```

Because the forward-slash is obviously already in use, as are many other typographical markers, I settled on the plus-sign (+). Words with alternations are transcribed with both symbols, separated by a plus sign. The left side of the plus sign is the standard vowel, and the right side is the Scots vowel.

```
all::DT/PDT/RB: { * oo l+w } :{all}:1558289
```

All the alternations except the MOUTH vowel are done in this way. If the rule does not apply, then the Scots symbol is simply deleted. If the rule does apply, both alternatives are put into different strings and scored appropriately. The rule has several environments that cover the different alternations. Some are quite specific, while some deal with all the alternations to a particular vowel, like |ei|. There is also a general environment that catches any alternations without specific environments. This makes it possible to assess the probability for any individual or group in a more fine-grained way than a simple all-in-one conversion would provide. The scores are shown in Table 4.10; some examples, in Table 4.11.

Environment	Score
V(r) → ei(r)	5
e[ei] → ai	4
V → i	3
V → V (general)	3
C → C	3

Table 4.10: Scores for alternation environments

'go' → 'gae'	#{ g * ou }# → #{ g * ei }#
'foot' → 'fit'	#{ f * uu ? }# → #{ f * i ? }#
'off' → 'aff'	#{ * o f }# → #{ * a f }#

Table 4.11: Alternations

Chapter 5

Voice Construction and Labeling

5.1 The Recorded Corpus

The recordings for the voice that I used to test my alterations to the lexicon were made in the Gorbals in October 2004 by a team of people led by Claire Timmins. They were recorded in a community hall. The recording is a spontaneous conversation between three Gorbals residents over the age of 60, two male and one female. Each speaker had his or her own microphone recording the voice on to a separate channel, and was in a room by him or herself; the participants could hear each other and the field worker through headphones. They were given a list of suggested topics but were otherwise speaking freely. The recordings were originally digitized at 24-bit, 48 kHz, but were converted to 16-bit encoding by the program `wavesurfer` in order to allow Festival and other Edinburgh Speech Tools to deal with the files; they were then downsampled to 16 kHz by Edinburgh Speech Tools `ch_wave` (Sjölander and Beskow 2005; Taylor et al. 1999, Mark Fraser, personal communication). The speaker used for this synthetic voice is a male, known by the pseudonym John.

The choice to use spontaneous speech was based on Fraser (2004), which compared two synthetic voices built from the same text, which in one case was produced spontaneously by the speaker and in one case was read by the speaker in the studio. The spontaneous speech was produced first in a two-hour session with the speaker, in which he discussed things of interest to him. The result was natural and informal and included slang and swearing. The studio speech was simply a reproduction of this speech by reading it. Evaluations suggested that the voice built from spontaneous speech sounded more natural. It is particularly sensible for this application, which attempts

to mimic spontaneous speech, to use this approach. It also avoids problems of imposing certain language on the speakers, allowing them to make their own linguistic choices at the phonological and lexical levels. Given what we know about Glaswegian speech and variation, this could be quite important in producing a natural sounding voice.

The recordings were manually segmented by Mark Fraser to find sections of speech that were relatively continuous and free from excessive disfluencies. This resulted in only about 11 minutes of continuous speech, which is a very small corpus. These segments were manually transcribed at the word level, including their disfluencies. I received the files as WAVE format sound files at 16-bit and the transcription as a Festival-format utterances file.

5.2 Labeling

During the process of voice building, the first thing that is done after segmentation and transcription of the sound files is to generate an initial labeling for them. This is done by Festival, taking the word-level transcriptions in the utterances file and creating an HTK format Master Label File containing initial phonetic transcriptions for all the files (Young, 1993). In order to do this, Festival uses its own version of the lexicon created by Unisyn for that accent; it does not use the running text transcription capability of Unisyn. Because there are disfluencies in the utterances, the lexicon that was used contains pronunciations for these disfluencies, which were added specifically to adapt the lexicon to what the speaker had actually said. Otherwise, the disfluencies could not have been labeled, and the labeling quality would have been very poor.

The lexicon was also tailored by adding certain spellings common in Scots and used in the initial transcriptions, even where the alternation capability of the lexicon would have been able to produce the correct pronunciation from the rules in the lexicon. Speaker-specific pronunciations of normal words were given a special spelling to distinguish them from the normal word, again following practice from the transcriptions, although normally the Unisyn exceptions list would be used for this. However, using the normal lexicon offers the advantage that both the normal and alternate pronunciation can be used. Some examples are shown below.

```
aboot::IN/RB/JJ: { @ . b * uu t } :{aboot}:0
awright::RB: { oo . t^ * ai t } :{awright}:0
eerhose::RB: { * ir r }.{ h ~ ou z } :{eer}{hose}:0
```

```
efter::IN: { * e f . t @r r } :{efter}:0
heid::NN/VB/VBP: { h * ii d } :{heid}:0
hink::VBP/VB/NN: { h * i ng k } :{hink}:0
oer::IN/RP/RB/NN: { * ou . @r r } :{oer}:0
ragirra::RB/NN: { tˆ @ . g * e . tˆ @r } :{ragirra}:0
thih::UH: { th * @ } :{thih}:0
wirrem::IN/JJ|PRP: { w * i }.{ tˆ @ m } :{wir}{rem}:0
```

'Aboot', 'efter', and 'heid' (about, after, head) are examples of words that can be produced by alternation, but instead were produced by the special spellings, to facilitate comparison with the voices built from the tailored Edinburgh Unisyn lexicon. 'Eerhose' is an example of the word that was added because the speaker used it, although it is not clear what word it is intended to be. The context, 'he was eerhose right clever', suggests that the word may be an odd pronunciation of 'always', but for such a small voice, it is easier to simply add the exceptions to a word list so they can be accurately labeled. Words such as 'hink', 'oer', 'wirrem', and 'ragirra' (think, over, with them, together) are all part of the variation of the dialect, but not variations that can yet be produced by the lexicon, because the rules involved may be lexically restricted and not yet been implemented.

'Wirrem' is a good example: it seems to show the operation of the rule converting /ð/ to tap in a word-initial environment, but it is possible that the collocation is quite strong and the environment acts almost like a word-medial environment, which would then fall under the aegis of that rule. The word 'awright' (all right) could be produced by rule, but the rule would also produce alternatives, and this is such a strong collocation that it is better to include it as one word. 'Thih' is a disfluency, as can be seen from its part of speech: 'UH', an interjection. These additions simply helped to adapt the lexicon to the speaker, for better results with such a small corpus.

The Glasgow output lexicon has multiple versions of each word, with the most probable one listed first. The probability for the different pronunciations of each word is computed in the same way as it is for running text, by determining the product of all the scores of all the alternatives taken in the word. Festival concatenates these words to create an initial transcription, which is then run through Festival's own postlexical processing. Although it is in theory possible for Festival to select between multiple alternatives, this requires the different alternatives to be labeled by some semantic or other criterion. This would not really be meaningful for these alternatives, and also

would not generate the desired statistical variation. Because this was not done, it will always choose the first alternative given in the lexicon. This means that although the improvement of the lexicon to match the Glaswegian accent will offer some improvement over using a lexicon for a different accent, neither the probabilistic nature of the accent nor the cross-word effects that can occur would be taken into account if this labeling were the final labeling, as in the usual procedure. During the standard labeling procedure, all that occurs after this step is a forced alignment of the chosen labels to the speech.

In the new labeling procedure, this is instead used as a baseline for initial training. Although it would also be possible to start with the transcription rated most probable by the Unisyn lexicon, I choose to begin with the transcription generated by Festival. This provides an analogous starting point for this voice as for the voices against which it will be compared (those built from the tailored Edinburgh lexicon).

Alignment and training are performed by HTK (Young, 1993), using the Viterbi algorithm for forced alignment to find the best time-aligned path through the given phones. During this step, HTK may make phone substitutions that are listed in a file, generally restricted to vowel reduction, to improve the labeling alignment. The phone models are then trained, using Baum–Welch re-estimation, with data from this alignment. A simple forced alignment and retraining is performed once more, followed by another forced alignment and training with mixture models up to eight mixtures. Finally, a forced alignment is performed once again, and the process normally terminates at this stage.

This process, however, allows no room for any of the possible running-text transcriptions from the lexicon to be checked as a possible transcription for the data. In order to do that, I have added an additional step. The Unisyn lexicon transformation is run on the source utterances, generating a file that contains all the possible options for each sentence. This file is then split up by sentence, and each version of each sentence is converted into HTK format and given an identifiable name indicating the number of the utterance it came from and where it fell in the order of possible sentences, so that each sentence can be uniquely identified. Each of these files is aligned with the speech, and the probability of its alignment recorded. The overall probability of the alignment is arrived at by summing the individual log probabilities of each phone match (there is no word-level match in the alignment process). The probabilities are compared, and the version with the highest probability is chosen as the new label for that file. The

phone models are then trained on the new labels, using the same procedure as before but without the step that introduces mixtures, and the alignment process is repeated, with the final best-aligned label being chosen as the true label and alignment.

5.2.1 Label Reconciliation for Utterance Building

After the labeling is done, the new aligned labels are incorporated into Festival's utterance structure as the final utterance structure is built. Festival is able to cope with some differences in the labeling generated by the alignment and its own generated labeling. These are normally confined to a different treatment of silences, the possible reduction of vowels, and the addition of closure labels before stops (a measure to improve the accuracy of the alignment). However, because I have used the Unisyn lexicon's transcriptions in labeling, the labels from the alignment procedure differed substantially from Festival's. This required the addition of new code to deal with possible label substitutions. New substitution rules were based on the names of the labels, because it was necessary to avoid any potential ambiguity. Allowable label substitutions were hardcoded into the utterance building function. These covered the substitutions attributable to differences in the lexicon, such as the occurrence of /ɾ/ for /r/. One of the less straightforward mechanisms included the deletion of /r/ where it had been deleted by the labeling software. This required a rule that operated after all the rest of the rules and deleted an /r/ where it corresponded to any other phone besides a tap, on the assumption that the corresponding /r/ had been deleted from the actual transcription. The substitution rules are shown in Table 5.1. Symbols to which the arrow points are the symbols retained. A double-headed arrow indicates that the substitution can occur in both directions.

Most of these substitution rules are due to Unisyn substitutions operating differently in the original transcription from the final transcription, but some require more explanation. In the case of the first rule, Festival sometimes indicates that the vowel has been reduced, but the labeling procedure labels as a full vowel. In this case, I assumed that Festival's production had been wrong and the labeling was correct.

Many of the rules, as the replacement of /f/ by /h/, only occur in one or two circumstances. In this case, the word 'something' was originally generated by Festival with the /f/ possibility, but the ultimate pronunciation used a /h/. This is likewise true for the replacement /z/ by /s/ and vice versa, which occurred in 'close' and 'houses', which may be pronounced with either option, depending on speaker or on part of speech.

Original ↔ Actual	**Comment**
@ → V	Festival performed a reduction that didn't occur
r → t^ r ↔ non-r	Unisyn r-rule
dh → t^	Unisyn dh/tap-rule
ou/uu/ai/ii ↔ ei a → ei our/ur → eir a → ae a ↔ o o → uh uu ↔ uh uu → i e ↔ i e → ii uh ↔ i oo → o iii → i	Vowel alternation rules
f → h	Pronunciation variation of "something"
hw ↔ w	Unisyn hw-loss
ow → uu	Unisyn MOUTH vowel rule
l → w	Unisyn consonant alternation
uuu → uu iii ↔ ii irr → ir	Different operation of the SVLR in changed environments
n! ← n	Restoration of lost syllabic n
f → th	Unisyn th/f-rule
? ↔ t	Unisyn glottal stop rule
z ↔ s	Alternate pronunciations of several words
s → k	Alternate pronunciations of "Celtic"

Table 5.1: Rules for label reconciliation during utterance building

The same is true of 'Celtic', which is pronounced with initial /s/ when referring to the football team, but with /k/ when referring to the culture. Depending on the other labels around the vowel, the SVLR may operate differently, and that is the reason for the substitutions of short vowels for long vowels or vice versa.

5.2.2 Correcting the Labeling

Because the automatic labeling procedure does not always produce good results, it generally requires some tuning. This is normally done by hand, and generally involves moving the labels to the correct location as determined on the spectrogram, with few or no changes in the actual labeling. Because of the small amount of data involved in this case, the automatic labeling does not function very well and this step is very important. Because this had already been done for the voice built from this data by Mark Fraser using the tailored Edinburgh lexicon, I elected not to do a true hand correction of the data using my labels, which would have been prohibitively time-consuming and would have duplicated much work already done. Instead, I wrote a script that checked the alignment of my labels with the hand corrected labels. The goal of this script was to maintain the labels that the new labeling procedure had chosen, while as much as possible using the label times from the hand-corrected data. To achieve this goal, the script used a dynamic programming alignment program called `dp`, part of Edinburgh Speech Tools, that performs dynamic programming alignment on label strings (Taylor et al., 1999). `dp` includes the ability to customize the cost of substitution between any pair of labels, with insertion and deletion defined as substitution of a label with a placeholder label. The costs that I used are shown in Table 5.2.

Event	Cost
Generic substitution	6
Generic deletion	7
Generic insertion	7
Insertion of `sp`	3
Insertion of `sil`	5
Deletion of a closure	2
Deletion of `sp`	3
Deletion of `sil`	5

Table 5.2: Costs for dynamic programming alignment

The customized costs contributed significantly to the accuracy of the alignment, which is absolutely essential for the time corrections done by the script to be accurate. Initial tests using HTK's dynamic programming alignment function, HResults, showed that some peculiarities of the labeling tended to cause problems. For example, the closure labels that are used in the labeling process were never present in the final labels, but it was not incorrect for them to be present, because Festival removes these during label reconciliation. Similarly, the silences were often different, and it was important to be able to specify that the insertion or deletion of a silence should not involve a heavy cost. The table shows that the costs for insertion and deletion of silences are small compared to the generic cost of insertion or deletion. Likewise, for a closure to be aligned with nothing (deleted) cost almost nothing, since it was more desirable that a closure should be mapped to nothing than that it should be mapped to another label that it would not match up with. (The order of presentation of the automatic labeling and the hand corrected labeling to `dp` made this a deletion rather than insertion, but this is actually arbitrary.) The low cost of substitutions relative to insertions and deletions is a result of needing to reconcile two slightly different labelings. Substitutions in this context are not as much of a problem as they normally are in speech recognition, where they indicate an error. Here, they only indicate that the label identities differed at that point.

Prior to running this alignment, I ran a separate script that deleted all silence labels with duration zero, including short interword silences. These silences were never present in the corresponding hand-corrected files, and thus could only provide a source of error. Wherever the number of labels was the same, I used the times provided by the hand correction. However, in many cases the number of labels differed, and so I could not take the times directly from the hand corrected labels.

There are two possible cases for differences in the number of labels.

In the first case, my labeling retained a label which was no longer in the hand corrected data. In these cases, if I had retained a short interword silence, the silence was simply deleted, and the resulting label sequence was the same as the hand corrected label sequence. These silences are normally inserted automatically during the labeling procedure, so the person who did the hand correction would know better whether one had actually occurred. If the extra label was not a silence, the start time for the extra label was the end time for the previously matched label, and its end time was its own end time from my labeling. For the following label, if it matched up with a label in the

hand corrected labeling, as was usually the case, its start time was the end time of the extra label, and its end time was the end time of the matching label. If it did not match up with the label in the hand corrected version, the process was repeated, its start time the end time of the previous label, and the start time from the next matching label as its end time. This result is shown in Table 5.3.

Auto label	x		y		z	
	t_1	t_2	t_2	t_3	t_3	t_4
Hand label	x				z	
	u_1	u_2			u_2	u_3
Result	x		y		z	
	u_1	u_2	u_2	t_3	t_3	u_3

Table 5.3: Automatic correction of labels when a label found in the automatic alignment is missing from the hand alignment

In other cases, I did not have a label which was in the hand corrected data. In that case, if I was missing a short interword pause, it was inserted, again because the hand-corrector would know more about these occurrences than the automatic procedure, and the result would be the same as the hand corrected labeling. Otherwise, the hand corrected label was deleted, and the time it previously occupied was assigned to the next label, so that the following matching label would span the entire time period. This outcome is shown in Table 5.4.

Auto label	x				z	
	t_1	t_2	t_2	t_3	t_3	t_4
Hand label	x		y		z	
	u_1	u_2	u_2	u_3	u_3	u_4
Result	x				z	
	u_1	u_2			u_2	u_4

Table 5.4: Automatic correction of labels when a label found in the hand alignment is missing from the automatic alignment

Unfortunately, this procedure is not ideal for every situation, because it is sometimes desirable to assign the time from the deleted phone to the previous phone. One common case of this is when a postvocalic /r/ is vocalized. The label correction perceives this as a deletion and assigns the time for the /r/ to the next phone, rather than the preceding

vowel, which is not correct. However, it is a straightforward procedure that generally functions well.

I will assess the effects of this process in the next chapter.

Chapter 6

Results

In assessing my results, I will be comparing four different voices. All of the voices use the same source sound files. They differ from each other on two dimensions. Two of the voices use the Glasgow Vernacular accent and will be referred to as the Gorbals voices. One of these used "hand-corrected labeling" (Gorbals, +LC) done by the script described in Section 5.2.2, and one had no label correction (Gorbals, -LC). Two of the voices use the Edinburgh accent with additions to the lexicon to tailor it to the voice and will be referred to as the Edinburgh voices. One of these has hand-corrected labeling (Edinburgh, +LC) and one has no label correction (Edinburgh, -LC). The latter two voices were created by Mark Fraser, who very kindly allowed me to use them for comparison.

6.1 Effects of the New Labeling Procedure

In order to assess the effects of the new labeling procedure, independent of its effect on the synthesis quality, I examined the labeling of 10 randomly chosen files from the 283 utterances from the Gorbals and Edinburgh -LC voices. I found that the different labeling procedure resulted in slightly different results, both in the time placement of labels and in which labels occurred. The following substitutions occurred in the Gorbals voice:

- An /r/ was deleted in the word 'years' (incorrect).
- The vowel in the word 'or' was substituted by a schwa (correct).

- An /ɹ/ was used instead of /ɾ/ in the words 'right' and 'regular' (incorrect). The speaker always uses taps in word-initial position.
- The full vowel /e/ was retained rather than reduced in the word 'tae' (correct).
- A /t/ was substituted for a glottal stop in the word 'out' (correct).
- An /r/ was deleted in the word 'hunters' (incorrect). The /r/ was not prominent, which might explain why it was missed.
- An /f/ was substituted for a /θ/ in the word 'three' (incorrect). The speaker categorically uses /θ/.
- In two cases, a /w/ was substituted for a /ʍ/ (both correct).
- An /r/ was deleted in the word 'after' (correct).
- A /t/ was substituted for a glottal stop in the word 'get' (incorrect).
- The full vowel /e/ failed to be substituted by a schwa in the word 'dae' (incorrect).
- A schwa was substituted for a full vowel in the word 'back' (incorrect).
- In the word 'Willie', the /ɪ/ was not replaced by /ʌ/ (incorrect).

Overall, there were 10 substitution mistakes and seven substitution improvements. In this selection of files, there were slightly more mistakes than improvements, but the numbers are fairly close, so it is not clear that there has been any improvement or decline overall in the accuracy. For the case of the word 'Willie', there was no rule to retract the /ɪ/, although it appears that there should have been (or perhaps an additional lexicon entry, as it may be lexically restricted). However, it is a positive result in that it confirms that some of the new substitutions are indeed necessary to accurately represent the accent. If it had not been possible to delete /r/s, the /r/ in the word 'after' could not have been deleted; likewise for the /t/ replacing a glottal stop.

It may be that with more training data, the accuracy of the flexible labeling procedure that I used would increase, because the system would be more able to distinguish between what in some cases are relatively subtle phonetic differences. Another possible problem is overtraining – because the system goes through two further iterations of training after the models have been fully trained using the normal procedure, and the amount of data is small, the models may be overtrained and therefore less accurate. It is also possible that if the initial labeling were better, the initial training would also

be better. The initial labeling used was the same as that used in the original labeling procedure, beginning with the concatenated Unisyn pronunciations of each word, rather than the running text transcriptions. Using this method affects which transcription for a word is most probable. For example, in lexicon mode every word appears to be followed by a pause, so final /r/ is more likely to be deleted. Because of this, it is likely that the utterances were trained with many of them missing, which would have affected the training data available for the category of /r/. This is the most probable reason for the mistakenly deleted /r/ is mentioned in the substitution list. A more desirable approach would be to train on the most likely transcription from the lexicon, although even that is not perfect, since it may be incorrect. A weighted training on all the possibilities would be quite interesting, and might also improve the results. Of course, hand-labeled training data would be the best; it might be possible to hand-label only a portion of the data to reduce the amount of time involved.

The same ambiguous results were found for the label timing. Manual examination allows label timing to be compared holistically, although not quantitatively. I found that the files tended to differ in regions—some regions were very similar, while others differed between the two sets of labels. On three files, the Gorbals labeling was better for more of the file, and on four the Edinburgh labeling was better. For the remaining files, the difference was not clear. Frequently the two labelings were very similar, and in particular they tended to be inaccurate in the same places. This is shown in Table 6.1, where the uncorrected labelings place `y` at 1.33 and 1.34 seconds, but the correct placement is at 1.26 seconds. Similarly, the glottal stop is actually located at 2.089 seconds, but the uncorrected labelings place it at 2.62 seconds. The uncorrected labelings differ at the third decimal place (by only thousandths of a second) in six cases, and of the cases where they differ at the first decimal place, all of the differences are by only hundredths of a second. So the change in labeling strategy does not seem to have necessarily made much change in the labeling.

A detailed quantitative comparison would require computation of which labels differed, and how much they differed by. This kind of comparison would certainly be desirable, but it is complex to make because of different labeling conventions introduced during the hand correction process.

Uncorrected Edinburgh		Uncorrected Gorbals		Corrected Edinburgh	
1.032	d_cl				
1.072	d	1.068	d	1.066000	d
1.208	a	1.204	a	1.206852	a
1.208	sp	1.204	sp		
1.334	y	1.342	y	1.265770	y
1.352	uu	1.352	uu	1.354000	uu
1.466	s	1.468	s	1.466000	s
1.476	t_cl				
1.532	t	1.484	t	1.486000	t
1.532	sp	1.484	sp		
1.538	t_cl				
1.544	t	1.49	t	1.498000	t
1.55	@	1.54	ei	1.528000	@
1.55	sp	1.54	sp		
1.588	d_cl				
1.616	d	1.632	d	1.624000	d
1.756	ei	1.792	@	1.778152	ei
1.756	sp	1.792	sp		
1.808	dh	1.804	dh	1.816386	dh
1.966	a	2.01	a	1.980000	a
2.062	Q_cl				
2.628	Q	2.622	Q	2.089342	?

Table 6.1: Labeling similarities and differences

6.2 Quality of the Gorbals Voice

The main evaluation of the new Glasgow voice is in comparison to the voice built using the modified Edinburgh lexicon. The purpose of this study is to determine whether creating a new accent that accurately reflects the accent of the voice material can improve the resulting synthetic voice. However, the overall quality of the voice is also important. There are therefore three sections of the evaluation: a subjective listening evaluation with comparisons to the Edinburgh voice, a listening test assessing the effect of accent and label differences on overall naturalness, and a small-scale test of the authenticity of the new voice as compared to the Edinburgh voice.

6.2.1 Subjective Evaluation

6.2.1.1 Overall Quality

Even within the general domain of sentences similar to those used to build the voice, the quality of the voice is somewhat inconsistent. Some synthesized utterances are highly intelligible and even very natural. For example, the sentence "I need a wee bit of advice" (`wee-advice.wav`) is easy to understand and very natural.[1] There is some noise, sounding somewhat like rustling, in the background, but this simply gives the effect of somebody speaking in an environment that is not perfectly quiet. Another high-quality sentence is "noo [now] see here lass, you cannae dae [do] that" (`lass-cannae.wav`). It is the kind of sentence that calls for a fair bit of vocal expression, and this is found in the synthesized result. The speaker sounds slightly indignant and very insistent. Interestingly, the transcription of the word 'you' is actually identical to the one achieved with 'ye', because the vernacular pronunciation is the first of the two alternatives for 'you' in the lexicon. In this case, it produces a result that sounds rather schwa-like. There are slight jumps in two places in the file, in the word 'here' and just before the word 'cannae', but no other noticeable lapses in quality. Other utterances have minor problems, such as "see my Danny, he's Celtic daft" (`celtic-daft.wav`). There is a pause after the word 'see' that is not exactly wrong, but sounds somewhat unusual. It is a result of the missing diphone `iii_m`. Notably, there is no realization of the /t/ in the word 'daft', even though it is in the transcription done by Festival. This is because the

[1] Files will be temporarily available at `http://www.ling.ed.ac.uk/~s0343949/thesis.html`. Filenames are given for reference; files are accessible by following links.

word is in the recordings—in fact the bigram "Celtic daft" occurs together. Otherwise, a less authentic pronunciation might be produced.

More commonly, there will be several problems with a given sentence. The sentence "I'll buy ye a drink, but I'm no a wine man" (`no-wine-man.wav`) has two clicks, near 'ye' and 'no'. There is also a slightly odd pause before the word 'man', which is the result of a long segment for the /n/ of the word 'wine' caused by prosodic lengthening in the source utterance. A sentence of similar quality is "what happens when you're working" (`happens-working.wav`), which also has an odd the pause before 'working'. This utterance also features an unnatural-sounding in the word 'happens', sounding higher and more front the normal /a/. Overall, it is a bit jumpy. Both of these files remain fairly intelligible, but they don't sound quite right.

Some utterances are quite bad, to the point where it affects not only their naturalness but also their intelligibility. The sentence "noo ye just don't gie [give] it a second look" (`second-look.wav`) has labeling so jumpy it almost sounds like someone gargling. There is a somewhat slurred sound between the first two words, and the prosody is quite uneven. Similarly, "those yins [ones] are nae good" (`yins-nae-good.wav`) also has prosodic problems, with the stress placed on 'our' rather than 'those', and it is generally jumpy. It also produces a pronunciation for 'nae' that sounds much more like 'no'. This mislabeling is discussed in Section 6.2.1.3.

6.2.1.2 Prosody

Prosody is overall of reasonable quality, but there are some sentences where it becomes the major problem of an otherwise good sentence. In the sentence "he worked in ra [the] pawn shop" (`pawn-shop.wav`), the prosody groups the sentence into sets of three words, with 'in' and 'shop' being particularly high relative to the preceding words (Figure 6.1). The prosodic pattern that would make most sense is to have peaks on 'worked' and 'pawn', and go down rather than up on the third words in each set. So both the overall pattern and the individual word matching are quite poor in this sentence. In a sentence "he couldnae get a job, errathing [everything] was closing up" (`job-errathing.wav`), the prosody is fine until the end, when it suddenly jumps up on 'up'. This is the most common problem with prosody. Another common problem is that the wrong word will receive stress, as in "those yins are nae good", discussed in the previous section.

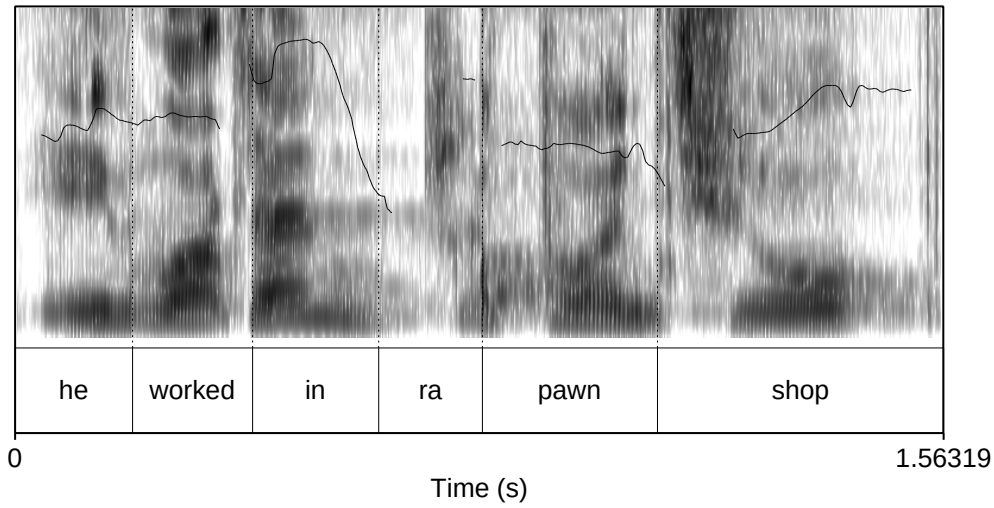

Figure 6.1: Spectrogram and pitch contour of "He worked in ra pawn shop."

6.2.1.3 Labeling

Problems with labeling sometimes cause bad units to be chosen, leading not just to the kind of jumpiness caused by time inaccuracies and slight label mismatches but to labels that actually sound wrong. This occurs in the sentence "if you go to a police they'll laugh you off" (`police-laugh.wav`). The final vowel in the word 'police' does not sound as a true /i/, a particular problem since the prosodic situation leads it to be stressed. The source label comes from the word 'really', which probably should not be labeled using the full high vowel, because it will frequently be unstressed and retracted. The unisyn lexicon has a rule that handles this, but the output for Scottish accents is the full vowel. This may work fine in many cases, but it did not work well here. To match this poor unit choice caused by bad labeling, the second part of the vowel was chosen from the word 'he', another likely candidate for being destressed. This may be a consequence of the general problem of just not having that many units, so that the choice of one or more nearby diphones may have been quite limited. The same thing does not occur when the word 'police' occurs by itself, so it appears to be a cascading effect of join and target costs. The two situations are compared in Figure 6.2.

A similar problem occurs in a sentence mentioned earlier, "those yins are nae good", where 'nae' sounds as 'no'. The labeling indicates that the source word was pronounced with a full vowel, but labeled with a schwa, and the word 'nae' was transcribed with

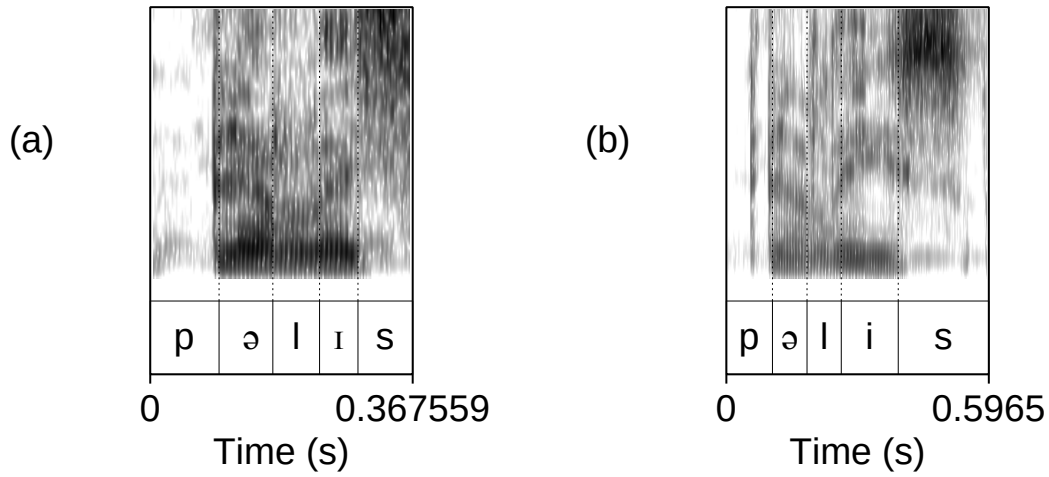

Figure 6.2: Spectrograms of "police" (a) embedded in a sentence (b) alone

a schwa, leading to a disconnect between the expected output and the actual output. The Festival transcription probably should not have used the destressed vowel at all, but it is particularly bad that having the destressed vowel led to getting a full vowel of the wrong type. Of course, there is natural variation between the two forms, but when specifying the dialect form, the normal form should not occur.

6.2.1.4 Missing Units

Because of the small size of the corpus, there are also several missing diphones. One was mentioned in Section 6.2.1.1, for the sentence "see my Danny, he's Celtic daft". In that case, the consequence is relatively mild—a pause that sounds slightly odd, but not really unnatural. On the other hand, a distinctly unnatural pause resulted from missing units in the sentence "is it no going to get any better" (`get-any-better.wav`), where the `?_e` diphone is missing, causing a pause between 'get' and 'any'. (The problem does not occur in the Edinburgh voice because of the labeling difference. In the source utterances, the sequence 'at every' was labeled with the full vowel beginning the word 'every', whereas in the Gorbals voice, this was labeled as reduced vowel. It is an interesting example of how the linear process of voice building causes problems to percolate from one level to another.) A more serious problem occurs in the sentence "he drives us hisself" (`drives-hisself.wav`), where the `v_z` diphone is missing, rendering that word almost unintelligible.

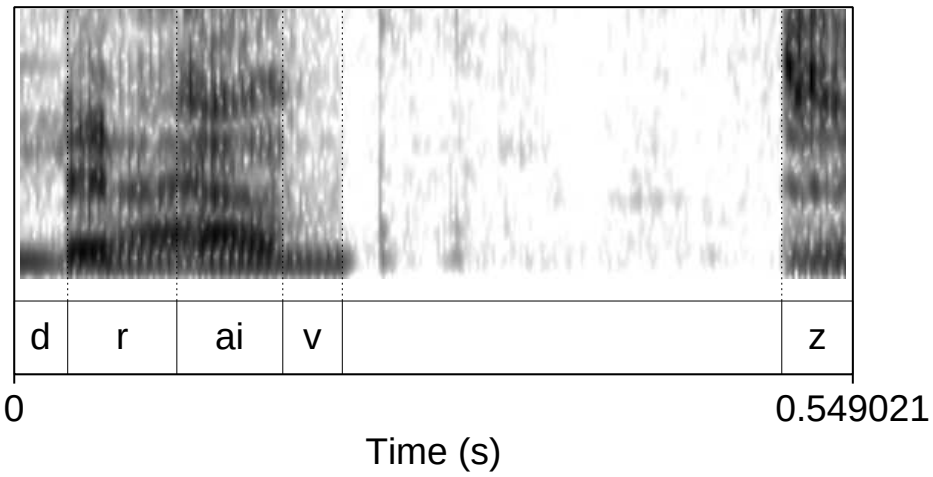

Figure 6.3: Spectrogram of "drives" with missing `v_z` diphone realized as a pause

6.2.1.5 Comparison with the Edinburgh Voice

There are two kinds of likely differences between the voices with straightforward sources. The first is segmental differences due to the labeling, where different labels have been chosen during the labeling process. These may come either from different choices by the automatic labeling or from different original pronunciations in the lexicons. The second is segmental differences due to the forms of the lexicons as they occur in Festival.

One example of the first type occurs in the sentence "see there was a crowd up in Easterhouse" (`see-easterhouse.wav`, `see-easterhouse-edi.wav`). Both lexicons present the diphthong /o^{w}/ rather than monophthong /ʉ/ in the final syllable of 'Easterhouse', but the sound of the Edinburgh voice is much more like /ʉ/, the desired result (like many place names, it retains its dialect form fairly strongly). The spectrograms of the two different realizations are shown in Figure 6.4. The Gorbals version does not look particularly like a diphthong in the spectrogram, but there is some formant movement, and it does sound as one. In this case, labeling all the diphthongs and monophthongs as diphthongs gives an advantage, because where a diphthong is given but the monophthong should occur, it's possible that the monophthong will occur. However, it does illustrate that these are perceptually separate categories, and should really be labeled separately and used in appropriate contexts. This is possible under the Gorbals lexicon, but doesn't occur in this case. Another example of this kind of difference is discussed

above in Section 6.2.1.4, where a labeling difference has caused the Gorbals voice to be missing a unit, so the same transcription is dealt with differently.

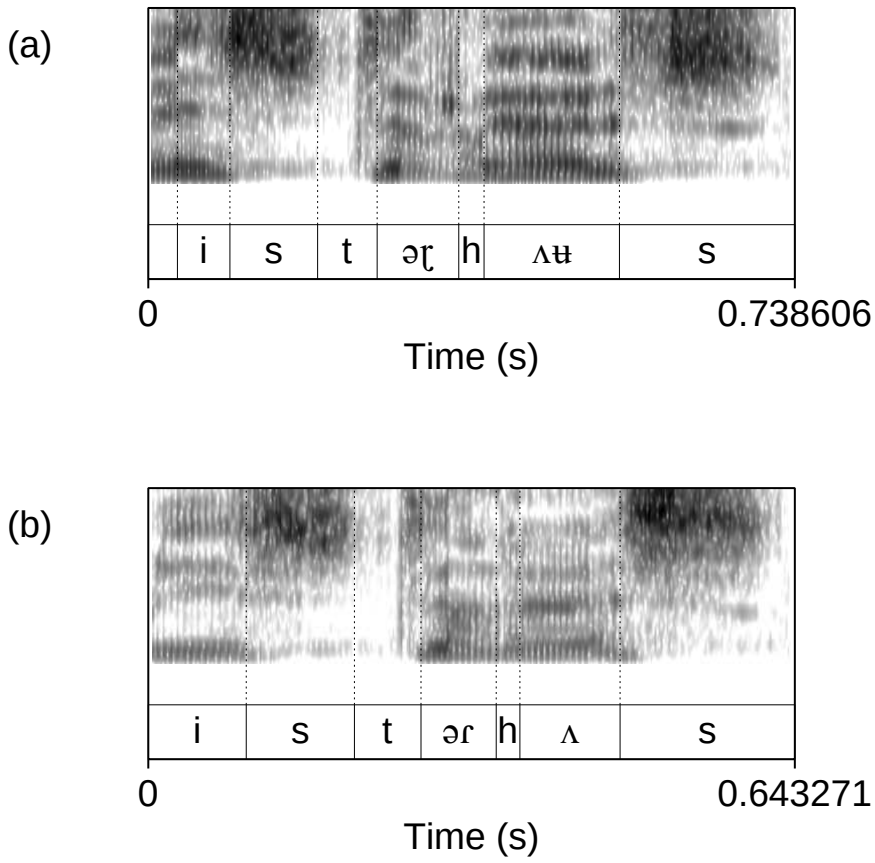

Figure 6.4: Spectrograms of "easterhouse" (a) Gorbals voice (b) Edinburgh voice

An example of the second type occurs in the phrase "he needs a wee bit of help" (`wee-help.wav`, `wee-help-edi.wav`). In this case, a particular adaptation of the modified Edinburgh lexicon is to provide an alternate pronunciation of the word 'of' that is reduced and does not include a final consonant (/ə/ vs /ɔv/). I failed to provide this alternate pronunciation because there was no means to do so without overwriting the normal pronunciation. As a result, Festival's transcription of the sentence is different for the two cases. The version from the Edinburgh voice sounds more natural, because this is in fact a good pronunciation of 'of' in unstressed positions, and the diphones chosen for the Gorbals version, which include the full vowel rather than a reduced vowel (although the reduced version is available), do not go well together. A less problematic example of this occurs in the phrase "you've just got to get on wirrit [with it]"

(`get-on-wirrit.wav`, `get-on-wirrit-edi.wav`). There is a slight artifact in the word 'wirrit' in both cases, but it is less noticeable in the Gorbals voice. The units do not come from the same files, because the Gorbals voice chooses a unit from the phrase 'fair enough' that is labeled as a tap in the Gorbals labeling but as an approximant in the Edinburgh labeling. The source for the unit in the Edinburgh labeling is the beginning of the word 'really', which is labeled with a long vowel in the Gorbals labeling, so it couldn't be used to join with the word 'it'.

There are also cases where the reason for the difference is less obvious. This is best exemplified by the sentence "he couldnae get a job, errathing was closing up" (`job-errathing-edi.wav`). There is a pause after the word 'job' that contains an artifact in the Gorbals voice but not in the Edinburgh voice (Figure 6.5). The units on the right side of the pause are different, and the source for the right side of the Gorbals pause comes from a short pause between words. The original short pause is not very quiet, so signal processing may have introduced this artifact from the slight noise that was present. Alternatively, the source for the pause before the `e` follows the word 'oors', so it may be colored by the /z/. The sound of the word 'a' is also slightly different, with something of a hiss present in the Gorbals version. The differing transcription here, with the full vowel for the word 'a', leads to a missing diphone, `ei_jh`, being called for, and replaced by two diphones with pauses in the middle. The sound of this pause may again not be very silent and may provide a kind of hissing artifact.

Another problem whose source is obvious, but whose different operation in a particular sentence is rather complex, is a mispronunciation of the word 'dae' in the Edinburgh voice for the sentence "noo see here lass, you cannae dae that" (`lass-cannae-edi.wav`). It is pronounced as the word 'die' would be. Investigation of the lexicon revealed that both the Gorbals and Edinburgh lexicons contain this as the correct pronunciation of this word. This is obviously not correct. However, for the Gorbals lexicon, the source of this word is from other occurrences of the same word, so the pronunciation is correct. In the Edinburgh lexicon, the sources are the word 'die' and the word 'like', both with the vowel indicated in the transcription. The reason for this appears to be that during hand correction, the hand corrector changed all occurrences of the incorrect vowel to the correct vowel, so that when looking for that vowel, instances of the word 'dae' are never found. However, the fundamental problem remained. Since I did not do true hand correction on my data, I avoided this problem.

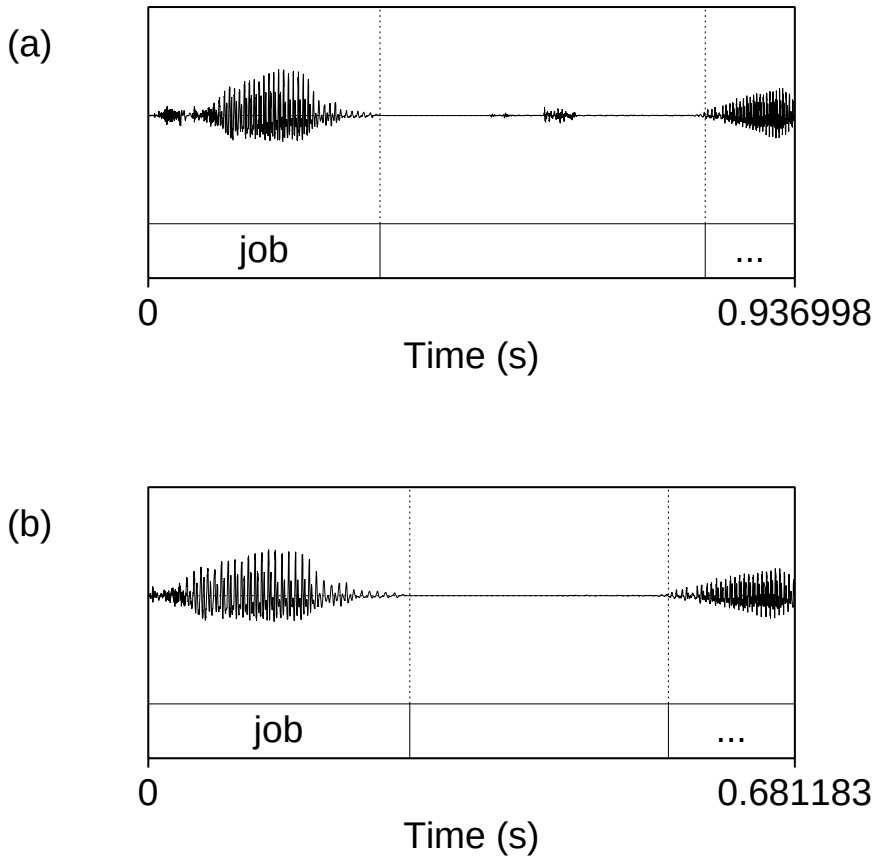

Figure 6.5: Waveforms of "job . . . " with artifact, (a) Gorbals voice (b) Edinburgh voice

Overall, the comparison with the Edinburgh voice indicates that both voices have problems. However, the Gorbals voice seems to be a step in the right direction in terms of getting and generating accurate labels for synthesis, and with a little more adaptation, it might well surpass the Edinburgh voice.

6.2.2 Listening Test for Naturalness

The listening test for naturalness was conducted using a forced-choice paradigm, making pairwise comparisons between all the voices. The forced-choice paradigm was chosen to avoid potential problems with interval scaling that can arise if the continuum on which the speech is being rated is not appropriate for the scale (Kreiman and Gerratt, 1998). Each pairwise comparison was separated into a different set; each set involved eight sentences; the same eight sentences were used in each set. I chose the eight sentences from a set that I had constructed manually for voice investigation. The sentences constructed for voice investigation were intended to explore various aspects of the voice using possible vernacular sentences based on the vocabulary contained in the source utterances. They were not vetted for possible lapses of the vernacular, but the restricted construction method, where they tended to be similar in style to the source sentences, probably prevented this from occurring. The sentences chosen for the test were reasonably intelligible and contained between six and ten words, an appropriate length for such a test. The sentences were:

1. I need a wee bit of advice.
2. Is it no going to get any better?
3. He couldnae get a job, errathing was closing up.
4. See my Danny, he's Celtic daft.
5. Rosie's gaun tae the shops the noo.
6. You've just got to get on wirrit.
7. Noo see here lass, you cannae dae that!
8. When I was wee I was a chimney sweep.

The listeners were instructed to listen to each pair of files and choose the file that sounded most natural to them. They were able to listen to the files as many times as they liked. It was recommended that they use headphones, but it was not required.

The test was conducted over the Internet.[2] Some effort was made to obscure which voice each file came from by instructing the browser's status bar not to display the name of the file when it was clicked or when the mouse was over it. Unfortunately, it is nearly impossible to entirely hide this information, because it can be accessed if the browser's file download box pops up. The arrangement of the files was randomized so that the files from one voice were not always presented first within a set, and so that the sentences were presented in a different order in each set. The files were rerandomized every several days during the period that the test was available, in order to make consultation between different listeners taking the test at different times difficult.

The test was designed to take less than half an hour, in order not to stretch the patience or capabilities of the listeners. This limited the number of sentences that could be used, since many comparisons needed to be made. It also made it inadvisable to check whether the listeners were being consistent in their judgments rather than random, because this would've cut down the number of sentences to a very small number and could have caused the results of the test to depend on the individual sentences used. However, the rerandomization of the files during the test period should have helped to cancel out any effect of inconsistency.

The listeners varied in geographical origin. There were a total of 24 participants, 19 native English speakers and five non-native. Of the 19 native English speakers, 10 were native speakers of North American English, of whom three are currently resident in the United Kingdom (two in Scotland and one in England). Thirteen of the participants used headphones and 11 used speakers. Nineteen of the participants reported experience listening to Scottish English, including all the non-native speakers. I was surprised by the high number of the participants who reported experience listening to Scottish English, given that my personal knowledge indicates that many of the North American speakers have never lived in the United Kingdom, and have only visited for short periods if at all. This illustrates the problem of asking questions that the participants can interpret in many ways. What constitutes experience for one person might not count for another. To ameliorate this problem, I could have used the phrasing "significant experience". This phrase still leaves it open to interpretation, but people with only a little experience might be less likely to answer yes to this question. All the non-native speakers, however, have lived in Scotland for at least several months, and so their answers for experience probably indicate reasonably significant experience, which

[2]The test will remain accessible temporarily at `http://www.ling.ed.ac.uk/~s0343949/gorbals_ab.html`, although submission has been disabled, in order to facilitate examination of the test.

is important because otherwise they might find it extremely difficult to understand the voices.

I hypothesized that the Gorbals voice would always be preferred over the Edinburgh voice, and that the comparison between the two Edinburgh voices would result in a preference for the voice with label correction, which amounts to hypothesizing a ranking of Gorbals, + LC > Gorbals, - LC > Edinburgh, + LC > Edinburgh, - LC. The results of this test do not support this hypothesis. Table 6.2 shows the results of the individual sets, where one voice was tested against another. If the column indicates zero for a particular set, that voice was not involved in that test set. None of the results were significant under a binomial sign test except the result for Set 3, which is significant at the level of p = 0.001. The two voices in this test differed in both accent and label correction. Trends in the other results indicate that label-corrected voices are generally preferred over voices without label correction, although this is not significant within the pairwise comparisons. There is no clear preference between the Gorbals voice and the Edinburgh voice, so it is not obvious why the comparison in Set 3 was significantly different while the comparison in Set 4 (the other comparison in which both factors differed) was not equally significant (although it is the comparison that came closest to significance).

The trends for the overall results, shown in Table 6.3, indicated that the label corrected voices are preferred more times than the non-label corrected voices. This is significant under a χ^2 test using the pairwise comparisons 1, 3, 4, and 6, with the total preferences for label corrected voices over uncorrected voices being 433 to 335, $\chi^2 = 12.505, p < 0.001$. The other two pairwise comparisons were not used because they were comparisons between voices with the same label correction status. Within each value of label correction, the Gorbals voice is preferred more times, although this effect is more slight even in the trends. A χ^2 test using pairwise comparisons 2, 3, 4, and 5 (those between voices with different accents) to test the overall effect of accent was not significant. It may be possible that with listeners using better quality equipment (more headphone users), and perhaps more listeners with experience with the kind of accents used, these results would achieve significance. Many listeners anecdotally reported difficulty in understanding what the person was saying, and may not have been able to identify some kinds of errors, such as pronunciation errors, that normally play a part in such evaluations. It was also common for listeners to report that they had difficulty distinguishing between the different versions of a file. Thus, the fact that they are not significant in this case indicates that there may not be a great quality change

achieved from simply using subtly different labeling. It is perhaps more important whether the change improves the authenticity of the accent from the standpoint of a listener familiar with the linguistic situation. This is discussed in the next section.

Set	**Gorbals, +LC**	**Gorbals, -LC**	**Edinburgh, +LC**	**Edinburgh, -LC**
1	104	88	0	0
2	97	0	95	0
3	120	0	0	72
4	0	86	106	0
5	0	91	0	101
6	0	0	103	89

Table 6.2: Pairwise results of the naturalness test

Voice	**Preferred**
Gorbals, +LC	321
Edinburgh, +LC	304
Gorbals, -LC	265
Edinburgh, -LC	262

Table 6.3: Overall results of the naturalness test

6.2.3 Listening Test for Authenticity

This test used the same paradigm as the previous test, but only two voices were compared: the Gorbals and Edinburgh voices with label correction. The other voices were not used for this comparison, because the difference in authenticity is only expected to be related to the change in accent. Because this test involved only one pairwise comparison, it was possible to use more sentences and also to test listener consistency. Twenty-one sentences were used for this test, including some of the sentences that were used in the test for naturalness. This was not a problem, because nobody did both tests. The sentences used were the following:

1. He needs a wee bit of help.
2. Is it no going to get any better?
3. He couldnae get a job, errathing was closing up.

4. See my Danny, he's Celtic daft.
5. Rosie's gaun tae the shops the noo.
6. You've just got to get on wirrit.
7. Noo see here lass, you cannae dae that!
8. When I was wee I was a chimney sweep.
9. You better not mess them aboot Danny.
10. He worked in ra pawn shop.
11. I hink it was his nephew that was drunk.
12. No he wasnae my pal, no it's just business.
13. It's hard and the money is no there.
14. We could go roond and sort him oot.
15. Don't go there, it's really awful.
16. Rosie's dressed up and aw, looking fancy.
17. I'll buy ye a drink, but I'm no a wine man.
18. See there was a crowd up in Easterhouse.
19. My bird, she loved it up there.
20. We go there awra time, he takes us hisself.
21. When your claes got mingin your ma would give ye a scolding.

The last sentence is 12 words rather than the usual six to ten, but it is a relatively quick and interesting sentence, and did not seem to present problems.

The test was divided into two sets, each containing the same 21 sentences. The sentences were presented in a different order within each set, and for each pair, the order in the second set was the reverse of the order in the first set. So if Sentence 1 had the sentence from the Gorbals voice on the left of the pair in the first set, Sentence 1 in the second set would have the sentence from the Edinburgh voice on the left. It would also appear in different places in the order of sentences. This enabled the testing of listener consistency by determining whether a listener who preferred the Gorbals voice for a sentence the first time it appeared would continue to do so the second time it appeared.

The task of determining whether the Gorbals voice is more authentically Glaswegian is likely to be most feasible for people with native-level familiarity with the Glaswegian accent. It is possible that some expert non-natives could do the task, but it would be difficult to determine whether such a person was genuinely able to do this. Thus, only Glaswegian natives were used as subjects for this test. Unfortunately, this author is not acquainted with many native Glaswegians, so it proved more difficult to find subjects for this test. As a result, only six listeners completed it. All the listeners had lived in Glasgow for at least 15 years. Two were currently resident in Glasgow, and the rest currently resident outwith Glasgow. I also asked the listeners to identify the area of the city that they most identified with. Three indicated Southside, one Parkhead, one the "Northwest" (further expanded as "George X/Maryhill"), and one a list of areas, beginning with Woodlands and Lenzie. Two had previous experience listening to synthesized voices. Four listeners used speakers, and two used headphones. I hypothesized that the listeners would find the Gorbals voice to be more authentic, although with such a small set of listeners I was not sure whether I would find anything significant.

The results are shown in Tables 6.4 and 6.5. The first table shows the results within each set. It is interesting to note that although each set contains the same sentences, the Edinburgh voice was more favored in the second set. This foreshadows the result shown for consistency in Table 6.6, namely that the listeners were not always consistent in their judgments. If the listeners were unable to judge between sentences and basically choosing randomly, we would expect to see roughly equal numbers of consistent and inconsistent choices. This was basically the case with Listeners 2, 3, and 4. Listeners 3 and 4 were hardly more consistent than chance, and Listener 2 was actually less consistent than chance would suggest (although not in any really meaningful way). Listener 6 was slightly more consistent. Only Listeners 1 and 5 were substantially consistent in their judgments.

The lack of consistency means that although the Edinburgh voice is slightly preferred over the Gorbals voice, the results of the test are probably not very meaningful. It seems that the voices are too similar even for native Glaswegians to make much difference between them. My hypothesis cannot be supported. However, one interesting result did come out of this test. There are three sentences in the test that showed large preferences one way or the other. The Gorbals version of Sentence 7, "noo see here lass, you cannae dae that", was always preferred. This is for the obvious reason that there is a mispronunciation in the Edinburgh voice for the sentence, as discussed in

Section 6.2.1.5. Two other sentences also had strong preferences, although in opposite directions. Sentence 8, "when I was wee, I was a chimney sweep" (`chimney-sweep.wav`, `chimney-sweep-edi.wav`), favored the Gorbals version by 11 preferences to one, and Sentence 10, "he worked in ra pawn shop" (`pawn-shop-edi.wav`), favored the Edinburgh version by 11 preferences to one. In both of these sentences, the primary difference is in the prosody. Sentence 10 is discussed in 6.2.1.2: the Gorbals version has quite an unnatural prosody. However, I am unable to distinguish an unnatural prosody in the case of Sentence 8. It seems plausible that in this case the native Glaswegians are able to perceive something non-Glaswegian about the prosody that is not accessible to me. This suggests that prosody plays an important part in the authenticity of a voice, and this might be a fruitful area for future research.

Set	**Gorbals**	**Edinburgh**
1	54	51
2	48	57

Table 6.4: Results of the authenticity test by set

Voice	**Preferred**
Gorbals	117
Edinburgh	135

Table 6.5: Overall results of the authenticity test

Listener	**Consistent**	**Inconsistent**
1	16	5
2	9	12
3	13	8
4	13	8
5	16	5
6	14	7

Table 6.6: Consistency of listeners in the authenticity test

Chapter 7

Conclusions and Future Work

7.1 Conclusions

The results discussed in the previous chapter are largely inconclusive. There was no clear improvement in the automatic labeling process, although some variation was more correctly captured by the Gorbals labeling. Subjective evaluation indicated that the voices were overall of similar quality: both voices produce some sentences very well, and other sentences not so well. Their particular problems are somewhat different, with both displaying artifacts and mispronunciations, but generally not in the same sentence. Obviously, the different labeling has some effect on the resulting voice, although the effects in individual sentences tend to cancel out, with some sentences being better in one voice and some on the other, giving no overall effect.

When asked to express pairwise preferences between the voices for naturalness, only one significant result was achieved, and that result seemed to involve the effect of label correction more than the difference in the accent, although the two are difficult to separate. Overall, the effect of label correction was significant, but this result is expected since label correction is known to improve the quality of a voice. The overall effect of accent was not significant in the preferences. Equally, native speakers were unable to distinguish one voice overall that more authentically reproduced the desired accent.

Given the inconclusiveness of these results, the question "What is the worth of this work?" must arise. The worth of this work lies primarily in the methods it develops for producing variation in speech synthesis. The development of probabilistic rules for the

Unisyn lexicon is a huge step forward in accurately representing the transcription not only of Glasgow Vernacular but also of any accent in which variation can occur: that is, all of them. Accents like Glasgow Vernacular benefit from this capacity on a phonetic and sociolinguistic level, but any accent could benefit from its potential capability to represent stylistic variation or differing speech rates. For example, more reduction occurs during fast, casual speech than during careful formal speech. By implementing reduction rules in a probabilistic fashion, this kind of variation could be easily represented. Once a rule has been implemented in this fashion, it is easy to turn it on and off, or adjust its degree of occurrence, by simply adjusting the probability allocated to the rule. Since the current implementation provides all possible alternatives, any desired alternative could be selected based on its score.

In addition to its greater ultimate accuracy and potential, this method is much more economical and generalizable than the method used in adapting the Edinburgh lexicon to produce the comparison voices. If the transcription of the source files at the word level had used standard words from the dictionary, rather than specially-created dialect words, the Edinburgh transcription would have been considerably less accurate. Although some labor-intensive addition of new words to the lexicon is inevitable with any significantly different dialect, the ability to produce many of the dialect words by simply adding vowel alternation to words already present in the lexicon could have considerably reduced the required labor. It also incorporates the work into the lexicon, rather than making it a special exception, thus allowing it to be the basis for future work. The adaptation from the Edinburgh lexicon also does not lay the groundwork for the later creation of variation at synthesis time, which would be a highly desirable future feature of the synthesis system.

The incorporation of complex processes into the lexicon also sheds light on the issue of greater incorporation of the lexicon system into the synthesis system. It is certainly valuable for lexicons to be able to stand apart from synthesis systems, and be used in many different systems. Likewise, it is useful to be able to plug multiple lexicons into a synthesis system depending on the needs of a particular situation. However, this work seems to suggest that it would be advantageous to synthesis systems if they could take advantage of more accurate transcriptions produced by more complex lexicon systems, most particularly if they were able to select among multiple transcriptions based on things like diphone availability and quality of units. Incorporating the method developed in this work could provide increased naturalness and quality, as well as allowing more accurate reproduction of accents—three major advantages.

7.2 Future Work

One major area of future work should be the greater understanding of Glasgow Vernacular on several levels, including the addition of a larger Scots vocabulary. In particular, the nature of vowel class membership remains slightly unclear, particularly the extent of its lexical penetration. Further fine-grained analysis of phonetic processes and their frequency would also be helpful, to build on the information already available in this area. This would support rules with a greater number of environments for more variation accuracy. It would also support the creation of a variety of voice profiles with different variation probabilities, which could be used to experiment with the sociolinguistic effects of different variation levels.

In terms of voice-building quality, it would be desirable to use a larger speech database and see if that improved the quality, since the overall quality of this voice is inconsistent. Labeling improvements would also help, especially altering the label process to begin with the most probable Unisyn labeling and avoiding overtraining by reducing the number of training iterations before the alternative labelings are evaluated. Improving the hand correction script to accurately reassign time from hand-corrected labels that do not appear in the automatic label alignment would be something quite specific to this voice, but might improve its quality and hope to assess the effect of label correction in general.

In the area of evaluation, a further field trial for authenticity is obviously desirable. The best situation would be to design a controlled testing situation with a large number of people in the Gorbals area, where the people are most qualified to judge whether the speech has been reproduced authentically. Such a large-scale trial would also support statistical significance testing, which could confirm whether any change is meaningful.

The most significant way to take this work forward would be to integrate the capability for variation into Festival. There are several possible ways that this could be done; I believe that most promising way is to use the Unisyn running text transcription abilities and select between the possibilities on a probabilistic basis, with the probability of an alternative determining how often it is selected. This would require integration between Festival's post-lexical processing and Unisyn's cross-word effects, resulting in a more integrated process with greater participation from the lexicon system. Speech synthesis will never sound truly natural until it is capable of all the variation produced

by human beings in natural speech, and now that the lexicon is capable of variation, this integration would be the next logical step on that very important road.

Bibliography

Aitken, A. J. (1984). Scots and English in Scotland. In Trudgill, P., editor, *Language in the British Isles*, pages 517–532. Cambridge University Press, Cambridge.

Black, A., Taylor, P., and Caley, R. (1999). The Festival Speech Synthesis System. http://www.cstr.ed.ac.uk/projects/festival/manual/festival_toc.html.

Fitt, S. (1997). The generation of regional pronunciations of English for speech synthesis. In *Proceedings of Eurospeech 97*, volume 5, pages 2447– 2450, Patras.

Fitt, S. (2000). *Documentation and User Guide to Unisyn Lexicon and Post-Lexical Rules*. Centre for Speech Technology Research, University of Edinburgh, Edinburgh.

Fitt, S. and Isard, S. (1998). Representing the environments for phonological processes in an accent-independent lexicon for synthesis of English. In *Proceedings of IcSLP 98*. Paper 0850.

Foulkes, P. and Docherty, G., editors (1999). *Urban Voices: Accent Studies in the British Isles*. Edward Arnold, London.

Fraser, M. (2004). Gorbals speech synthesis. Master's thesis, Division of Informatics, University of Edinburgh.

Hunt, A. and Black, A. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *Proceedings of ICASSP 96*, pages 373–376, Atlanta, GA.

Johnston, P. (1997). Regional variation. In Jones, C., editor, *The Edinburgh History of the Scots Language*, pages 433–513. Edinburgh University Press, Edinburgh.

Jones, C. (2002). *The English Language in Scotland: An Introduction to Scots*. Tuckwell Press, East Linton, East Lothian.

Kreiman, J. and Gerratt, B. (1998). Validity of rating scale measurements of voice quality. *Journal of the Acoustical Society of America*, 104(3 Pt 1):1598–1608.

Macafee, C. (1994). *Traditional Dialect in the Modern World: A Glasgow Case Study*. Lang, Frankfurt am Main.

Macaulay, R. (1978). Variation and consistency in Glaswegian English. In Trudgill (1978), pages 132–143.

Macaulay, R. and Trevelyan, G. (1977). *Language, Social Class, and Education: A Glasgow Study*. Edinburgh University Press, Edinburgh.

Mather, J. Y. and Speitel, H. H. (1986). *Linguistic Atlas of Scotland*, volume 1–3. Croon Helm, London.

Romaine, S. (1978). Postvocalic /r/ in Scottish English: Sound change in progress? In Trudgill (1978), pages 144–157.

Scobbie, J., Hewlett, N., and Turk, A. (1999). Standard English in Edinburgh and Glasgow: the Scottish Vowel Length Rule revealed. In Foulkes and Docherty (1999), pages 230–245.

Sjölander, K. and Beskow, J. (2005). Wavesurfer. http://www.speech.kth.se/wavesurfer/.

Stuart-Smith, J. (1999a). Glasgow: accent and voice quality. In Foulkes and Docherty (1999), pages 203–222.

Stuart-Smith, J. (1999b). Glottals past and present: A study of T-glottaling in Glaswegian. *Leeds Studies in English*, 30:181–204.

Stuart-Smith, J. (2003). The phonology of Modern Urban Scots. In Corbett, J., McClure, J. D., and Stuart-Smith, J., editors, *The Edinburgh Companion to Scots*, pages 110–137. Edinburgh University Press, Edinburgh.

Stuart-Smith, J. and Tweedie, F. (2000). Accent change in Glaswegian: A sociophonetic investigation. Technical report, Final Report to the Leverhulme Trust (Grant no. F/179/AX). http://www.arts.gla.ac.uk/SESLL/EngLang/research/accent/accent1.htm.

Taylor, P., Caley, R., Black, A., and King, S. (1999). The Edinburgh Speech Tools library. http://www.cstr.ed.ac.uk/projects/speechtools.html.

Timmins, C., Tweedie, F., and Stuart-Smith, J. (2004). Accent change in Glaswegian (1997 corpus): Results for consonant variables. http://www.arts.gla.ac.uk/SESLL/EngLang/research/accent/results070904.pdf.

Trudgill, P., editor (1978). *Sociolinguistic Patterns in British English*. Edward Arnold, London.

Wells, J. (1982). *Accents of English*, volume 1 and 2. Cambridge University Press, Cambridge.

Williams, B. and Isard, S. (1997). A keyvowel approach to the synthesis of regional accents of English. In *Proceedings of Eurospeech 97*, volume 5, pages 2435–2438, Patras.

Young, S. J. (1993). The HTK Hidden Markov Model Toolkit: Design and philosophy. Technical report, Cambridge University Engineering Department, CUED/F-INFENG/TR.152, Cambridge.